TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

^	Opérateur d'estimation.
$'$	Opérateur de transposé.
$\{\mathbf{a}_t\}_{t=1..T}$	Suite des T vecteurs acoustiques.
$\{\mathbf{x}_i\}_{i=1..N}$	Ensemble de N i-vecteurs.
$< . >$	Espérance mathématique.
a	Vecteur acoustique de type MFCC.
B	Matrice de covariance interclasse (*Between Classes Covariance matrix*).
BCCN	Normalisation avec la matrice de covariance interclasses (*Between Classes Covariance Normalization*).
C	Nombre de composantes gaussiennes d'un modèle GMM.
D	Dimension des i-vecteurs.
DCT	Transformée en cosinus discrète.
DER	Erreur de la structuration (*Diarization Error Rate*)
e	i-vecteur d'enrôlement.
EER	Erreur de la vérification (*Equal Error rate*)
F	Dimension des vecteurs acoustiques.
$f(\mathbf{x})$	Fonction de densité de probabilité.
FFT	Transformée de Fourier rapide.
GMM	Modèle de Mélange de Gaussiennes (*Gaussian Mixture Model*).
h	Largeur de bande.
JFA	Analyse conjointe de facteurs (*Joint Factor Analysis*).
LDA	Analyse discriminante linéaire (*Linear Discriminant Analysis*).
LPC	*Linear Prediction Coefficients*.
MFCC	*Mel Frequency Cepstral Coefficients*.
M_l	Modèle du locuteur l.
MS	Algorithme de décalage de la moyenne (*Mean Shift*).
p	Facteur d'élagage.
PCA	Analyse en composantes principales (*Principal Component Analysis*).

PLDA	Analyse discriminante linéaire probabiliste (*Probabilistic Linear Discriminant Analysis*).
PLP	*Perceptual Linear Predictive*.
PPCA	Analyse en composantes principales probabiliste (*Probabilistic Principal Component Analysis*).
r	Enregistrement audio.
RBM	Machine de Boltzmann restreinte.
SAC	Similarité angulaire de cosinus.
SSR	Stratégie sélective de regroupement.
STR	Stratégie totale de regroupement.
$\mathbf{t}$	i-vecteur du test.
WCCN	Normalisation avec la matrice de covariance intraclasse (*Within Class Covariance Normalization*).
$\mathbf{x}$	i-vecteur.
η	Quantité de la variabilité retenue pour la PCA.

INTRODUCTION

Dans le contexte technologique actuel où le monde est réduit à un petit village quasi connecté, plusieurs nouveaux besoins ont émergé et sont même devenus vitaux. Nous citons, à guise d'exemple, le besoin d'accéder et de gérer à distance en toute sécurité ses comptes bancaires ou bien ses bases de données hautement confidentielles, à travers son téléphone intelligent. Ce genre de tâches serait sans doute plus accessible si on peut les accomplir de façon orale ; autrement dit, avec un recours minimal à des outils tels qu'un clavier ou une souris.

Outre qu'un moyen naturel de communication entre la plupart des êtres humains, la parole peut également servir à caractériser l'identité du locuteur en se basant sur les traits physiologiques et comportementaux véhiculés par le signal vocal. Ceci a donné naissance à une discipline quasi contemporaine dénommée la reconnaissance automatique du locuteur. À vrai dire, la reconnaissance du locuteur est l'une des sous-disciplines du traitement de la parole. Elle se divise à son tour en plusieurs sous domaines, soit la vérification, l'identification et le regroupement en locuteurs. En effet, la reconnaissance du locuteur n'est certainement pas le moyen biométrique le moins intrusif ni le plus fiable qui soit. Or, lorsque la voix d'une personne est la seule caractéristique physique dont on dispose, la vérification biométrique de son identité n'est possible qu'à travers la discipline de la reconnaissance du locuteur.

Les deux dernières décennies ont été largement marquées par plusieurs progrès en termes de performances et de robustesse des systèmes de la reconnaissance du locuteur. Ceci est dû principalement à l'adoption des recherches en reconnaissance du locuteur par l'institut américain (*National Institute of Standards and Technology*, NIST). Le tout récent progrès, voire même le plus important, est la proposition de la représentation de la parole par des vecteurs à faible dimension, dénommés les i-vecteurs (Dehak, *et al.*, 2009)(Dehak, *et al.*, 2011b). De par sa dimensionnalité modérée (typiquement dans les centaines) et sa richesse en matière d'information utile, l'espace des i-vecteurs a énormément aidé les chercheurs en reconnaissance du locuteur, et même en d'autres domaines connexes, pour ce qui est de la

conception des classificateurs simples et hautement robustes. Dans le contexte de cet espace de représentation, les classificateurs génératifs, tels que l'analyse discriminante linéaire probabiliste (PLDA) (Kenny, 2010a), ont dominé le domaine de la reconnaissance du locuteur. Cependant, de simples classificateurs à base de la similarité angulaire du cosinus (SAC) (Dehak, *et al.*, 2010)(Senoussaoui, *et al.*, 2013a) restent également compétitifs.

❖ **Objectifs**

De nos jours, les recherches en reconnaissance du locuteur et notamment en vérification du locuteur ont atteint un niveau de maturité très avancé en ce qui concerne les taux de reconnaissance. Dans certaines tâches assez difficiles de la vérification, telles que définies par NIST, l'erreur de la vérification est aux alentours de 2 %. Ainsi, l'idée d'entamer une piste de recherche dans le but d'améliorer ces taux d'erreur semble peu prometteuse.

D'une manière générale, notre principal objectif dans cette thèse est de rendre les systèmes actuels de reconnaissance du locuteur plus robustes face aux facteurs dégradants, afin de s'en servir dans des milieux réels, en dehors des environnements contrôlés des laboratoires. Plus particulièrement, nous nous intéressons uniquement aux systèmes de reconnaissance du locuteur implémentés dans le contexte de l'espace des i-vecteurs.

Dans la première partie de cette thèse, nous entamons deux facteurs limitant la robustesse des systèmes de vérification. Le premier facteur à traiter est celui de la dépendance du type du canal de transmission et/ou d'enregistrement (téléphone/microphone). Quant au deuxième facteur, il est lié à la différence physiologique entre les voix des locuteurs et celles des locutrices. Ainsi, nos objectifs dans cette partie sont de rendre ces systèmes indépendants du type du canal et du genre du locuteur. Il est important de souligner que jusqu'à présent, lors de ses campagnes d'évaluation de la reconnaissance du locuteur (*Speaker Recognition Evaluation*, SRE), l'institut américain NIST contourne généralement ces deux facteurs en imposant certaines conditions et en fournissant même des informations supplémentaires.

Dans la deuxième partie de la thèse, nous avons élargi nos horizons de recherche en abordant un autre sous-domaine de la reconnaissance du locuteur, à savoir, le regroupement en

locuteurs des segments vocaux. En effet, nous gardons toujours notre contexte de la représentation de la parole dans l'espace des i-vecteurs.

❖ **Méthodologies**

Afin d'aboutir à notre objectif de rendre les systèmes de vérification du locuteur indépendants du canal et du genre, nous considérons les systèmes de l'état de l'art (c.-à-d. les systèmes à base du PLDA et ceux à base de la SAC) comme des systèmes de base dans toutes nos séries d'expériences.

Quant aux recherches de la deuxième partie, notre algorithme de base sera l'algorithme de *décalage de la moyenne* (Mean Shift, MS) (Fukunaga, et al., 1975). Le *Décalage de la moyenne* est un algorithme non paramétrique de recherche des modes de distributions de probabilité inconnues. Dans le cadre de cette thèse, nous proposons une nouvelle version de cet algorithme à base de la distance angulaire du cosinus.

❖ **Organisation de la thèse**

La thèse est organisée en huit chapitres de la façon suivante. Dans le Chapitre 1, nous présenterons une mise en contexte afin de faciliter la compréhension aux lecteurs non familiers avec le domaine de la reconnaissance du locuteur. Dans le Chapitre 2, la définition de l'espace des i-vecteurs ainsi que la procédure détaillée de leurs extractions seront fournies. Le Chapitre 3 est dédié à la présentation des systèmes de vérification du locuteur basés sur les deux méthodes de la classification citées ci-dessus (le PLDA et la SAC). Les Chapitres 4 et 5 seront respectivement consacrés aux questions de l'indépendance du canal et celle du genre des systèmes de la vérification du locuteur. Le Chapitre 6 détaillera l'algorithme de base de décalage de la moyenne (MS) ainsi que la nouvelle version de cet algorithme proposée dans cette thèse. Ensuite, dans les Chapitres 7 et 8, nous étudierons l'efficacité de notre nouvelle version de MS à base de la distance angulaire du cosinus respectivement face à la tâche de regroupement en locuteurs des grandes bases de données (*Speaker Clustering*) ainsi qu'à celle de la structuration des flux audio en tours de parole (*Diarization*). L'Annexe I fournit une preuve mathématique de convergence de la nouvelle version de l'algorithme de

Décalage de la moyenne. Finalement, l'Annexe II fournit des illustrations graphiques montrant les effets de l'application successive des méthodes de la normalisation des i-vecteurs sur la structuration en tours de parole via le *Décalage de la moyenne*.

CHAPITRE 1

RECONNAISSANCE DU LOCUTEUR

1.1 Généralités

La reconnaissance du locuteur est la tâche d'authentification de l'identité d'une personne en analysant uniquement sa voix. La reconnaissance du locuteur est une activité qui exploite les techniques du traitement automatique de la parole et de la reconnaissance des formes pour accomplir le but ultime de la biométrie, à savoir, l'authentification des personnes. Outre que la reconnaissance du locuteur, la discipline du traitement automatique de la parole recouvre également une vaste panoplie d'activités (voir Figure 1.1), tels que : le codage, la compression, la synthèse et la reconnaissance de la parole, la reconnaissance de la langue et du dialecte, la reconnaissance des émotions et l'estimation de l'âge, etc. Tout comme les autres activités de la reconnaissance automatique de la parole, la reconnaissance du locuteur est quasi contemporaine ; sa première apparition remonte au début des années 1960 (Furui, 2005). Cependant, cet axe de recherche est devenu très actif notamment dans les deux dernières décennies où le besoin de ce genre d'applications ne cesse de croitre. Les champs d'application de la reconnaissance du locuteur sont assez larges et mosaïques allant des applications de nature individuelle et domestique jusqu'à la sécurité en passant par l'indexation, la structuration et la recherche dans des bases de données multimédia. Il faut noter que la voix n'est pas d'une nature purement biométrique, raison pour laquelle son application reste limitée pour des fins juridiques (Bimbot, *et al.*, 2005).

À son tour, la reconnaissance du locuteur se divise essentiellement en quatre axes de recherche (voir Figure 1.1), à savoir, la vérification du locuteur, l'identification du locuteur, le regroupement en locuteurs (*Speaker Clustering*) et la structuration en tours de parole (Speaker Diarization).

En présence d'un segment vocal, la vérification du locuteur est la procédure d'authentification de l'identité d'une personne prétendant être l'émetteur (le locuteur) de ce

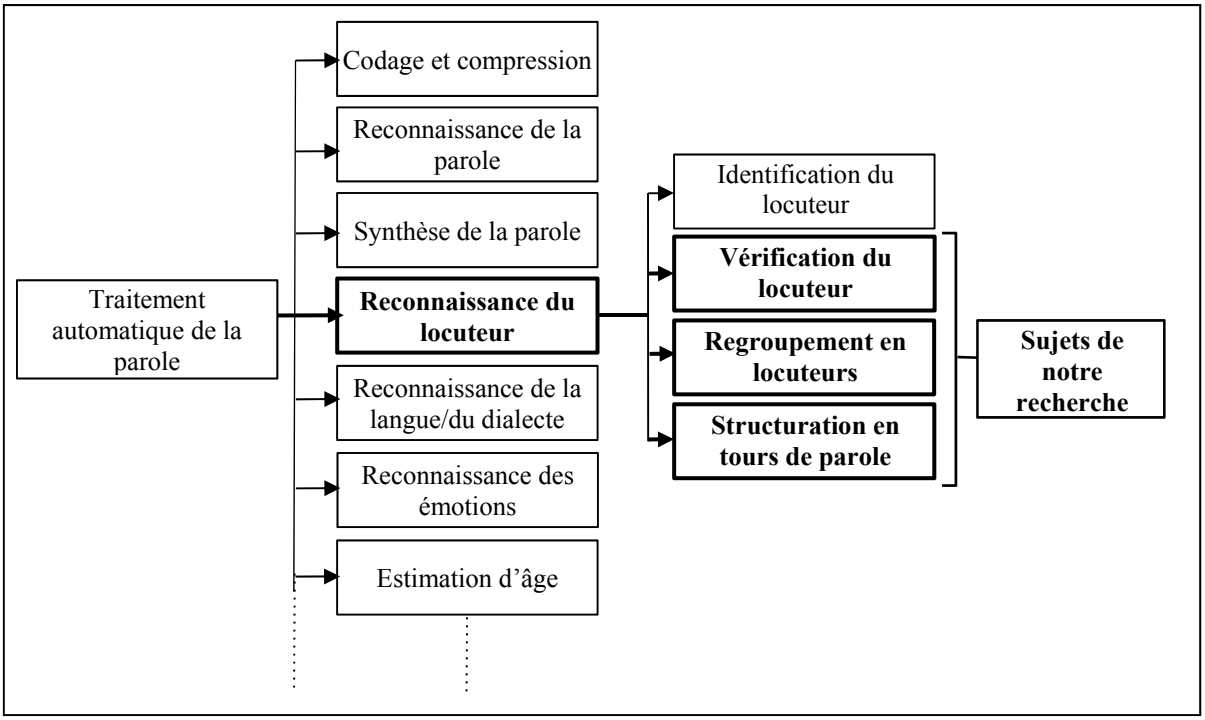

Figure 1.1 Arbre du traitement automatique de la parole avec une focalisation (en Gras) sur les sujets traités dans cette thèse.

segment (Campbell, 1997)(Furui, 2008)(Kinnunen, *et al.*, 2010). Pour pouvoir accomplir cette tâche de vérification, le système doit avoir préalablement construit un modèle de la parole du locuteur en question, c'est ce qu'on nomme dans le jargon de la reconnaissance du locuteur la phase d'enrôlement d'un locuteur (voir Figure 1.2). Bien que le concept de la vérification du locuteur paraisse simple, l'implémentation d'un tel système est assez complexe. La communauté de recherche rassemblée par l'institut américain NIST[1] a accordé plus d'intérêt à la vérification du locuteur qu'au reste des tâches de la reconnaissance automatique du locuteur (voir ces tâches sur la Figure 1.1) de fait qu'elle en constitue l'unité de base.

L'identification du locuteur, quant à elle, consiste à sélectionner l'identité la plus probable parmi un ensemble déjà connu par le système et l'attribuer à un segment vocal (Campbell, 1997)(Furui, 2008)(Kinnunen, *et al.*, 2010). L'identification peut être vue également comme

[1] National Institute of Standards and Technology : http://www.itl.nist.gov/iad/mig/tests/sre/

une succession d'exécutions de la tâche de vérification du locuteur entre le segment en question et l'ensemble des locuteurs clients déjà enrôlés par le système. Dans le cas des systèmes d'identification dits à ensemble ouvert (open-set systems), la réponse du système peut être aussi un rejet d'identité si ce dernier a conclu que la voix du segment de parole ne correspond à aucun de ses clients.

L'identification et la vérification du locuteur peuvent être accomplies toutes les deux dans un mode dépendant ou indépendant du texte (Campbell, 1997)(Furui, 2008)(Kinnunen, *et al.*, 2010). Dans le premier mode, le locuteur est prié d'être coopératif dans le sens où il doit prononcer un texte dicté par le système (texte prompt) ou bien le fixer antérieurement, par exemple un mot de passe. Quant au mode indépendant du texte, le contenu linguistique du segment vocal n'est pas imposé et il est souvent considéré comme une source supplémentaire de variabilité indésirable pour la distinction des locuteurs.

La classification et le regroupement automatiques des données non étiquetées pour déterminer leur structure intrinsèque est un problème traditionnel dans le domaine de la reconnaissance des formes et l'apprentissage-machine. Étant donné un corpus de données de segments vocaux où chaque segment est prononcé par un seul locuteur et chaque locuteur a au moins un segment, le processus du regroupement de cette base de données en considérant le locuteur comme étant la classe d'intérêt est nommé le regroupement en locuteurs (*Speaker Clustering*)(Kotti, *et al.*, 2008)(Van Leeuwen, 2010)(Senoussaoui, *et al.*, 2013b).

Le problème du regroupement en locuteurs est souvent confondu avec le problème de la structuration en tours de parole (*Speaker Diarization*). Contrairement au problème du regroupement en locuteurs qui cherche à regrouper des segments vocaux enregistrés sur différents fichiers, la structuration en tours de parole consiste à segmenter et regrouper un seul flux audio en régions homogènes correspondantes à la parole des locuteurs participants. Généralement la structuration en tours de parole est caractérisée par sa capacité de répondre aux questions : « Qui parle ? Quand ?» (Tranter, *et al.*, 2006).

Les recherches abordées dans cette thèse se focalisent sur trois des quatre problèmes de la reconnaissance du locuteur: la vérification du locuteur, le regroupement en locuteurs et la

structuration en tours de paroles. Dans ce chapitre, nous détaillerons davantage ces trois tâches tout en nous appuyant sur la vérification du locuteur vu son importance pour le reste des tâches.

1.1.1 Types de la variabilité de la parole

Avant d'entamer les sections qui fournissent plus de détails sur les sous-disciplines de la reconnaissance du locuteur, il est important d'introduire les types de variabilités de la parole. En plus du contenu linguistique, le signal vocal véhicule aussi des informations sur : l'identité du locuteur, son genre, son état émotionnel, son état de santé, sa tranche d'âge, etc. Dans une base de données de plusieurs locuteurs où chaque locuteur a au moins un enregistrement[2], les enregistrements d'un même locuteur peuvent être effectués dans différents lieux (par exemple : au bureau, à la maison, à l'extérieur, dans une chambre insonorisée, etc.) et en utilisant différents types de canaux (par exemple : téléphone fixe, téléphone portable, microphone, etc.). Usuellement, nous pouvons limiter toute cette gamme d'informations ou de variabilités dans ces trois catégories suivantes :

- *La variabilité* ***intrasession*** : il s'agit de la variabilité perçue au cours d'un même enregistrement et qui peut être causée par un changement de l'état émotionnel, de lieu, d'emplacement par rapport au microphone ou bien du microphone lui-même, etc.
- *La variabilité* ***intersessions*** : il s'agit de la variabilité perçue entre l'ensemble des enregistrements d'un même locuteur, elle est ainsi due aux mêmes causes que l'intrasession, mais elle est susceptible d'être plus importante vu les différences possibles entre les dates et les circonstances d'enregistrement.
- *La variabilité* ***interlocuteurs*** : il s'agit de la variabilité perçue entre les distincts locuteurs et due principalement à la différence d'anatomie des cordes vocales, ça pourrait être aussi due au genre, à l'âge, à la langue, etc.

[2] Dans le jargon du traitement de la parole, le mot *session* est aussi utilisé pour référer à un enregistrement.

Selon l'application du traitement de la parole, l'utilité/nuisibilité de ces variabilités restera toujours à déterminer. Par exemple, les traits caractérisant un locuteur sont assez importants et utiles pour une application de la reconnaissance du locuteur en même temps qu'ils sont très nuisibles à une application de reconnaissance de la parole ou de la langue. Il est à noter que dans le cas de la vérification du locuteur, la variabilité interlocuteur constitue l'unique information utile pour la distinction des locuteurs. De plus, toute variabilité outre que l'interlocuteur est considérée comme nuisible et souvent nommée *effet du canal*.

1.2 Vérification du locuteur

1.2.1 Définition et utilité de la tâche

La vérification du locuteur consiste à examiner, d'une façon automatique, la véracité de l'hypothèse qui stipule qu'un segment vocal est émis par un locuteur préalablement enrôlé par le système (voir Figure 1.2) et habituellement nommé un locuteur cible ou un locuteur client (Campbell, 1997)(Furui, 2008)(Kinnunen, *et al.*, 2010). En d'autres termes, la vérification du locuteur consiste à mesurer la ressemblance entre deux segments de parole afin de juger s'ils appartiennent ou non au même locuteur. Dans le cadre de cette thèse, nous nous intéressons particulièrement au problème de la vérification du locuteur en mode indépendant du texte. Cette vérification en mode indépendant du texte est certainement beaucoup plus complexe du fait que le contenu phonétique génère une source additionnelle de variabilité indésirable pour la séparation entre les différents locuteurs.

1.2.2 Extraction des vecteurs de caractéristiques à court terme

Le signal de la parole sous sa forme analogique est très complexe, mais également non stationnaire ce qui rend sa manipulation assez difficile. Une phase de représentation de ce signal sous forme d'une succession de vecteurs de valeurs numériques (nommés trames) est indispensable pour la majorité des applications du traitement automatique de la parole. Cette phase est souvent appelée le paramétrage du signal ou bien la phase d'extraction des vecteurs de caractéristiques à court terme. Afin de pouvoir réaliser ce traitement, le signal vocal est

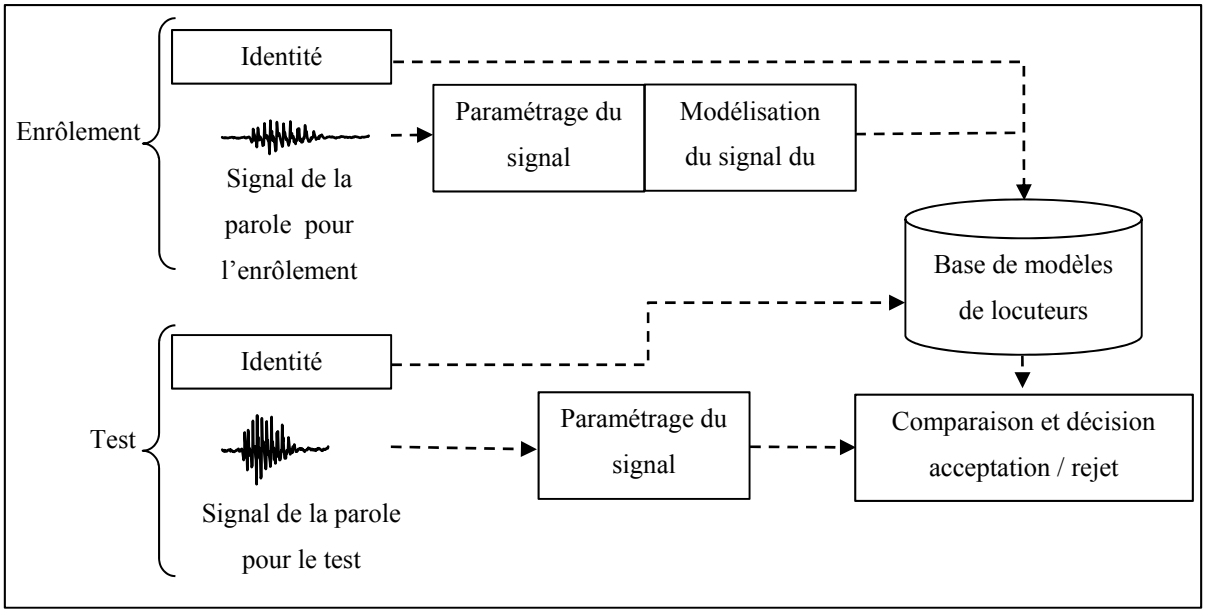

Figure 1.2 Schéma typique d'un système de vérification du locuteur : phase d'enrôlement et phase de test.

supposé être stationnaire pendant des intervalles de temps de très courte durée mesurée en milliseconde. En nous basant sur cette hypothèse, nous pouvons extraire un ensemble des caractéristiques qui représentent ce signal le plus fidèlement possible, à l'intérieur d'une fenêtre glissante (avec une période typique d'avancement égale à 10 *ms*) d'une fenêtre temporelle de très courte durée (typiquement 25 *ms*). Cette représentation a l'avantage de bien modéliser l'information spectrale du signal vocal tout en conservant sa nature temporelle. Cependant, la représentation à court terme est caractérisée par sa durée variable due au fait que le nombre des trames (c.-à-d. les vecteurs de caractéristiques) extraites est relatif à la durée réelle du signal. Ce caractère variable de la durée des données complexifie davantage la tâche de la classification.

Parmi ces vecteurs de caractéristiques à court terme qui modélisent les informations spectrales du conduit vocal, nous citons les Mel Frequency Cepstral Coefficients (MFCC), les Perceptual Linear Predictive (PLP) et les Linear Prediction Coefficients (LPC) qui sont fréquemment utilisés dans les applications de la reconnaissance de la parole. Étant donné que la représentation MFCC à court terme est la plus répandue dans le domaine de la vérification

du locuteur (Kinnunen, *et al.*, 2010), les MFCC seront les vecteurs de caractéristiques adoptés pour la réalisation des recherches présentées dans cette thèse.

1.2.3 Modélisation

Au fil des années, plusieurs approches ont été étudiées dans le domaine de la vérification du locuteur afin de modéliser la structure complexe qui caractérise la voix d'un locuteur donné. En premiers temps, on a adopté les approches basées sur le concept de « *templates matching* », par exemple la quantification vectorielle (QV) (Soong et al., 1985) et la programmation dynamique (Dynamic Time Warping, DTW) (Furui, 1981). Par la suite, les modèles génératifs provoquent un vrai engouement dans ce domaine, entre autres les modèles de Markov cachés (Hidden Markov Models, HMM) dans le cas des systèmes dépendants du texte (BenZeghiba, *et al.*, 2003) et les modèles de mélanges de gaussiennes (Gaussian Mixture Models, GMM) dans le cas des systèmes indépendants du texte (Reynolds, 1992)(Reynolds, 1995)(Reynolds, *et al.*, 2000b). Parallèlement avec tout ce qui est génératif, les modèles discriminants tels que les réseaux de neurones (RN) (Farrell, *et al.*, 1994) et les machines à vecteur de support (Support Vector Machine, SVM) ont également trouvé leur place dans ce domaine (Campbell, 2006). Récemment, l'introduction des Machines de Boltzmann a été observée (Stafylakis, *et al*., 2012a)(Senoussaoui, *et al*., 2012)(Stafylakis, *et al*., 2012b)(Vasilakakis, *et al*., 2013).

1.2.3.1 Modèles de mélanges de gaussiennes

La majorité des systèmes considérés actuellement comme l'état de l'art dans le domaine de la vérification du locuteur en mode indépendant du texte sont basés d'une manière ou d'une autre dans leurs structures internes sur les GMM. Du point de vue théorique, un GMM avec un nombre important de composantes (gaussiennes) est capable de modéliser une vaste gamme de distributions de probabilité, quelle que soit la complexité. Dans l'espace acoustique, les vecteurs de caractéristiques à court terme sont principalement répartis en fonction des classes phonétiques et de la nature du son émis (son voisé/son non voisé) ce qui favorise des estimateurs multimodaux, tels qu'un GMM, pour modéliser leur distribution. En

outre, les modèles génératifs comme les HMM et les GMM ont l'avantage de pouvoir gérer le problème de la variabilité de durée engendrée par la représentation à court terme.

Soit $S = \{\mathbf{a}_t\}_{t=1..T}$ une suite de T vecteurs acoustiques (MFCC par exemple) représentant le signal vocal d'un locuteur l et soit M_l le modèle GMM composé de C composantes gaussiennes modélisant un locuteur l. La vraisemblance d'un vecteur acoustique de la séquence S alignée avec le modèle M_l est donnée par la formule suivante :

$$\mathrm{P}(\mathbf{a}_t \mid M_l) = \sum_{c=1}^{C} \pi_c^l N(\mathbf{a}_t, \mu_c^l, \Sigma_c^l) \tag{1.1}$$

où π_c^l est le poids de la $c^{\text{ième}}$ composante du mélange et $N(.,\mu,\Sigma)$ est la fonction de densité de probabilité d'une loi normale multidimensionnelle de vecteur moyen μ et de matrice de covariance Σ donnée par :

$$N(\mathbf{a},\mu,\Sigma) = \frac{1}{2\pi^{F/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{a}-\mu)'\Sigma^{-1}(\mathbf{a}-\mu)\right] \tag{1.2}$$

où F est la dimension des vecteurs acoustiques, l'apostrophe est un opérateur de transposée et le symbole $|.|$ représente le déterminant d'une matrice. De ce fait, un modèle GMM d'un locuteur l est caractérisé par $M_l = \{\pi_c^l, \mu_c^l, \Sigma_c^l\}_{c=1..C}$.

L'apprentissage d'un GMM selon le principe du maximum de vraisemblance (Maximum Likelihood, ML) consiste tout simplement à estimer les paramètres $\{\pi_c^l, \mu_c^l, \Sigma_c^l\}_{c=1..C}$ qui maximisent la vraisemblance totale $p(S \mid M_l)$ de la suite S des vecteurs d'apprentissage calculée de la façon suivante :

$$\begin{aligned} \mathrm{P}(S \mid M_l) &= \prod_{t=1}^{T} \mathrm{P}(\mathbf{a}_t \mid M_l) \\ &= \prod_{t=1}^{T} \sum_{c=1}^{C} \pi_c^l N(\mathbf{a}_t, \mu_c^l, \Sigma_c^l) \end{aligned} \tag{1.3}$$

Une solution analytique du problème de la maximisation de fonction de vraisemblance n'est pas évidente. Cependant, un algorithme bien connu sous l'acronyme EM (Expectation

Maximization) (Dempster, *et al.*, 1977) a apporté une solution itérative à ce problème. Les formules de réestimation itérative des paramètres $\{\pi_c^l, \mu_c^l, \Sigma_c^l\}_{c=1..C}$ d'un modèle du locuteur *l* via l'algorithme EM sont les suivantes :

$$\pi_c^{l,(i)} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{P}\left(c \mid \mathbf{a}_t, M_l^{(i-1)}\right) \tag{1.4}$$

$$\mu_c^{l,(i)} = \frac{\sum_{t=1}^{T} \mathrm{P}\left(c \mid \mathbf{a}_t, M_l^{(i-1)}\right)\mathbf{a}_t}{\sum_{t=1}^{T} \mathrm{P}\left(c \mid \mathbf{a}_t, M_l^{(i-1)}\right)} \tag{1.5}$$

$$\Sigma_c^{l,(i)} = \frac{\sum_{t=1}^{T} \mathrm{P}\left(c \mid \mathbf{a}_t, M_l^{(i-1)}\right)\mathbf{a}_t^2}{\sum_{t=1}^{T} \mathrm{P}\left(c \mid \mathbf{a}_t, M_l^{(i-1)}\right)} - \mu_c^{l,(i)} \tag{1.6}$$

où *i* est un indice de l'itération en cours et *i*-1 est celui de l'itération précédente et la probabilité *a posteriori* $\mathrm{P}\left(c \mid \mathbf{a}_t, M_l^{(i-1)}\right)$ est calculée comme suit :

$$\mathrm{P}\left(c \mid \mathbf{a}_t, M_l^{(i-1)}\right) = \frac{\pi_c^{l,(i-1)} N(\mathbf{a}_t, \mu_c^{l,(i-1)}, \Sigma_c^{l,(i-1)})}{\sum_{c=1}^{C} \pi_c^{l,(i-1)} N(\mathbf{a}_t, \mu_c^{l,(i-1)}, \Sigma_c^{l,(i-1)})} \tag{1.7}$$

Dans la pratique, un nombre important des composantes GMM, typiquement entre 256 et 2048, est essentiel pour modéliser la voix d'un locuteur. Ainsi une quantité de données relativement grande est nécessaire afin d'aboutir à une estimation suffisamment robuste du modèle du locuteur. Pour pallier à ce problème, deux solutions sont proposées : 1) assumer l'indépendance entre les composantes des vecteurs de caractéristiques, ce qui permet d'estimer des matrices de covariances diagonales. Cependant, cette solution a l'inconvénient de dégrader les performances du modèle. 2) l'utilisation d'un modèle très bien entrainé en utilisant une grande quantité de différents locuteurs comme un modèle *a priori* nommé le modèle du monde[3] (Universal Background Model, UBM) (Reynolds, *et al.*, 2000b); à la présence d'une faible quantité de données d'un locuteur *l*, la mise à jour (l'adaptation) des paramètres d'UBM en utilisant ces données permet la création d'un modèle suffisamment robuste. En fait, cette architecture est très connue sous l'appellation GMM-UBM. La

[3] Nous utilisons la lettre grec oméga en majuscule (Ω) pour référer au modèle du monde UBM.

méthode d'adaptation la plus répandue dans la littérature est une méthode bayésienne connue sous l'appellation MAP (*Maximum A Posteriori*) (Gauvain, *et al.*, 1994) (Reynolds, *et al.*, 2000b). Les étapes d'adaptation itérative des paramètres du modèle du monde $\Omega = \{\pi_c^{\Omega}, \mu_c^{\Omega}, \Sigma_c^{\Omega}\}_{c=1..C}$ aux données d'apprentissage du locuteur *l* selon le critère MAP sont :

- Pour chaque instant *t* et chaque gaussienne *c* du modèle Ω nous calculons la probabilité *a posteriori* :

$$P(c \mid \mathbf{a}_t, \Omega) = \frac{\pi_c^{\Omega} N(\mathbf{a}_t, \mu_c^{\Omega}, \Sigma_c^{\Omega})}{\sum_{c=1}^{C} \pi_c^{\Omega} N(\mathbf{a}_t, \mu_c^{\Omega}, \Sigma_c^{\Omega})} \tag{1.8}$$

- Nous utilisons ces probabilités *a posteriori* pour calculer respectivement des statistiques d'ordre zéro, un et deux comme suit :

$$P(c \mid \Omega) = \sum_{t=1}^{T} P(c \mid \mathbf{a}_t, \Omega) \tag{1.9}$$

$$\mathrm{E}_c[\mathbf{a}] = \frac{1}{P(c \mid \Omega)} \sum_{t=1}^{T} P(c \mid \mathbf{a}_t, \Omega) \mathbf{a}_t \tag{1.10}$$

$$\mathrm{E}_c[\mathbf{a}\mathbf{a}'] = \frac{1}{P(c \mid \Omega)} \sum_{t=1}^{T} P(c \mid \mathbf{a}_t, \Omega) \mathbf{a}_t \mathbf{a}_t' \tag{1.11}$$

- Enfin, ces statistiques servent à estimer les paramètres adaptés :

$$\pi_c^l = \left(\alpha_c^{\pi} \frac{P(c \mid \Omega)}{T} + (1 - \alpha_c^{\pi}) \pi_c^{\Omega} \right) \gamma \tag{1.12}$$

$$\mu_c^l = \alpha_c^{\mu} \mathrm{E}_c[\mathbf{a}] + (1 - \alpha_c^{\mu}) \mu_c^{\Omega} \tag{1.13}$$

$$\Sigma_c^l = \alpha_c^{\Sigma} \mathrm{E}_c[\mathbf{a}\,\mathbf{a}']_c + (1 - \alpha_c^{\Sigma})(\Sigma_c^{\Omega} + \mu_c^{\Omega} \mu_c'^{\Omega}) - \mu_c^{\Omega} \mu_c'^{\Omega} \tag{1.14}$$

où le paramètre γ assure que la sommation des poids estimés de toutes les gaussiennes du mélange soit égale à un, le facteur de pondération α sert à régulariser les paramètres du modèle adapté (c.-à-d. modèle du locuteur) en fonction de l'importance accordée aux

paramètres *a priori* (c.-à-d. les paramètres du modèle du monde Ω). Autrement dit, un $\alpha = 1$ implique que les nouveaux paramètres sont estimés selon le critère de maximum de vraisemblance. Ainsi, ce facteur est estimé en fonction de la quantité des données alignées avec la composante *c* et un hyper-paramètre τ comme suit :

$$\alpha_c = \frac{P(c|\Omega)}{P(c|\Omega) + \tau} \tag{1.15}$$

En observant les formules d'adaptation (1.12, 1.13 et 1.14) et la formule de pondération (1.15), il s'avère clair que seules les gaussiennes du modèle du monde recevant une quantité significative de données d'adaptation seront adaptées. Dans la pratique, seule l'adaptation des vecteurs des moyennes prouve son efficacité (Reynolds, *et al.*, 2000b).

Le modèle de mélanges est devenu l'état de l'art dans le domaine de la reconnaissance du locuteur de par sa capacité à modéliser les distributions amplement complexes des vecteurs acoustiques. Cependant, un modèle GMM compte toujours sur d'autres méthodes pour contrer les variabilités nuisibles qui dégradent la qualité du modèle. Ainsi, plusieurs méthodes ont été proposées dans la littérature pour combler ces lacunes, tantôt dans l'espace acoustique (la normalisation de la moyenne et la variance des vecteurs acoustiques), tantôt au niveau des scores de ressemblance calculés souvent entre deux segments vocaux (*t-norm*, *z-norm* etc.). En addition à tout ceci, une nouvelle allure a été donnée au modèle GMM en le représentant seulement via la concaténation des vecteurs moyens de ses composantes gaussiennes. En fait, ce vecteur est largement connu sous l'appellation de *supervecteur*.

1.2.3.2 Modèle d'analyse conjointe de facteurs (JFA)

Le modèle d'analyse conjointe de facteurs proposé par (Kenny *et al.,* 2008), (Joint Factor Analysis, JFA), est un modèle génératif des variables cachées qui tente principalement de s'attaquer au problème de la disparité entre les données d'enrôlement et celles de test dans l'espace des supervecteurs des GMM. L'hypothèse de base de ce modèle stipule qu'un supervecteur **S** de dimension *CF* dépendant du locuteur et du canal (c.-à-d. un supervecteur

estimé d'un segment donné d'un locuteur donné) peut se diviser en deux supervecteurs statistiquement indépendants comme suit :

$$\mathbf{S} = \mathbf{s} + \mathbf{c} \tag{1.16}$$

où **s** est le supervecteur qui dépend uniquement du locuteur et **c** est celui qui dépend du canal. Le supervecteur **s** dépendant du locuteur est donné par la formule suivante :

$$\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} \tag{1.17}$$

où les supervecteurs moyens **m** représentés souvent par le supervecteur du modèle du monde UBM, **V** connue par la matrice des voix propres, c'est une matrice rectangulaire de dimension $CF \times R_s$ (avec $R_s \ll CF$), elle peut aussi être vue comme une matrice de projection en dimension réduite, elle modélise principalement la variabilité interlocuteur. **D** est une matrice diagonale de dimension $CF \times CF$ qui joue un rôle similaire à celui de l'information *a priori* dans l'adaptation MAP du modèle GMM-UBM. Les vecteurs cachés **y** et **z** sont respectivement les facteurs du locuteur (*speaker factors*) et les facteurs communs (*common factors*). Les vecteurs cachés suivent une distribution normale standard $N(.,\mathbf{0},\mathbf{I})$ où **0** est un vecteur moyen dont toutes ses composantes sont égales à zéro et **I** est une matrice-identité. Donc, le supervecteur **s** suit une distribution gaussienne d'un vecteur moyen **m** et d'une matrice de covariance qui est égale à $\mathbf{V}\mathbf{V}' + \mathbf{D}^2$.

Le supervecteur **c** dépendant du canal est donné par la formule suivante :

$$\mathbf{c} = \mathbf{U}\mathbf{x} \tag{1.18}$$

où la matrice **U**, dite matrice des canaux propres (*eigenchannels*), est une matrice rectangulaire de dimension $CF \times R_c$ (avec $R_c \ll CF$), ses colonnes constituent les axes de l'espace du canal de dimension réduite. Le vecteur caché **x** de dimension R_c distribué selon une loi normale standard est appelé le vecteur des facteurs du canal (*channel factors*). Ainsi, le supervecteur **c** suit une loi normale d'un vecteur moyen nul et d'une matrice de covariance égale à $\mathbf{U}\mathbf{U}'$.

De plus amples détails sur l'implémentation du modèle JFA se trouvent dans (Kenny, *et al.* 2008).

1.2.3.3 Espace des i-vecteurs

Le succès de l'architecture GMM-UBM dans le domaine de la reconnaissance du locuteur ainsi que la proposition des supervecteurs ont donné naissance au modèle JFA. Ce modèle tente de remédier au problème des disparités causées par les effets du canal dans l'espace de très grande dimension des supervecteurs. À son tour, le modèle JFA a largement préparé le terrain à l'arrivée de la représentation de la parole par des vecteurs à faible dimension dénommés par le terme d'i-vecteurs (Dehak, *et al.*, 2009)(Dehak, *et al.*, 2011b). La définition la plus élémentaire d'un i-vecteur est de le considérer comme étant la projection d'un supervecteur dans un espace de dimension réduite. Cet espace, à l'inverse des deux sous-espaces de JFA, ne fait aucune distinction entre les différents types de variabilité (c.-à-d. la variabilité interlocuteur et celle due au canal), ainsi, il est nommé l'espace de la variabilité totale.

Mathématiquement, un supervecteur **S** d'un segment audio dépendant du locuteur et du canal peut se réécrire sous la forme suivante :

$$\mathbf{S} = \mathbf{m} + \mathbf{T}\mathbf{x} \tag{1.19}$$

où le supervecteur moyen **m** de dimension CF est le supervecteur de modèle du monde UBM, la matrice rectangulaire **T** de dimension $CF \times D$ (avec $D \ll CF$) est nommée la matrice de la variabilité totale (c.-à-d. la variabilité interlocuteur et celle due au canal) et **x** est un vecteur caché de dimension D qui suit une distribution normale standard. Ainsi, le supervecteur **S** suit une distribution normale du vecteur moyen **m** et de la matrice de covariance qui est égale à $\mathbf{T}\mathbf{T}'$. Le modèle génératif proposé à (1.19) peut être estimé via un modèle d'analyse en facteurs ou bien via une analyse en composantes principales probabiliste (*Probalistic Principal Component Analysis*, PPCA) (Dehak, *et al.*, 2011b)(Kenny,

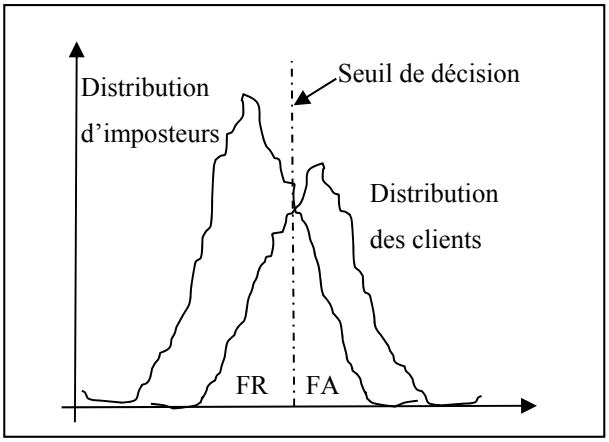

Figure 1.3 Schéma approximatif des distributions des scores des locuteurs clients et imposteurs ainsi que le seuil de décision optimale.

2012)(Bishop, 2007). Il est à noter qu'une estimation ponctuelle (*point estimate*) du vecteur caché **x** nous donne le i-vecteur.

De par sa nature, l'espace de la variabilité totale (c.-à-d. l'espace des i-vecteurs) contient simultanément la variabilité désirable et indésirable à la distinction des locuteurs. Ainsi un mécanisme de filtration est nécessaire pour avoir un système de reconnaissance du locuteur robuste. Dés l'apparition de l'espace des i-vecteurs dans le domaine de la vérification du locuteur, deux méthodes de classification sont émergées dans ce domaine, à savoir l'analyse discriminante linéaire probabiliste (*Probabilistic Linear Discriminant Analysis*, PLDA) comme un modèle génératif (Prince, *et al.*, 2007)(Kenny, 2010a) et la similarité angulaire du cosinus (SAC) comme un modèle à base de distance/similarité (Dehak, *et al.*, 2010)(Dehak, *et al.*, 2011b)(Senoussaoui, *et al.*, 2013a).

1.2.3.4 Analyse discriminante linéaire probabiliste (PLDA)

Contrairement à ce qu'indique son appellation[4] , le PLDA est un modèle purement génératif, il n'est en réalité qu'une version miniature du modèle JFA (voir la section 1.2.3.2) opérant

[4] Cette appellation est une analogie avec celle de PPCA qui est une version probabiliste de PCA ordinaire (Bishop, 2007).

dans le sous-espace des i-vecteurs au lieu de celui des supervecteurs. Le modèle de base de PLDA est apparu pour la première fois dans le domaine de l'imagerie où il est utilisé pour la reconnaissance faciale (Prince, *et al.*, 2007). Il est ensuite introduit dans la vérification du locuteur avec deux variantes et quelques modifications dans l'algorithme d'apprentissage (Kenny, 2010a). L'apprentissage et la vérification via ce modèle seront détaillés plus tard au Chapitre 3 de cette thèse.

1.2.3.5 Similarité angulaire

La similarité angulaire à base de cosinus repose sur le calcul de l'angle entre deux vecteurs afin de mesurer leur similarité tout en ignorant leurs amplitudes. Dans l'espace de la variabilité totale (c.-à-d. l'espace des i-vecteurs), cette métrique a été appliquée avec succès au problème de la vérification du locuteur (Dehak, *et al.*, 2010)(Dehak, *et al.*, 2011b)(Senoussaoui, *et al.*, 2013a) ainsi qu'à d'autres problèmes, tels que la reconnaissance du langage (Dehak, *et al.*, 2011c) et la structuration en tours de parole (Shum, *et al.*, 2011)(Shum, *et al.*, 2012)(Shum, *et al.*, 2013), etc. Plus de détails sur la compensation des effets du canal dans le cadre de la similarité angulaire seront fournis ultérieurement au Chapitre 3.

1.2.3.6 Machines de Boltzmann

Une Machine de Boltzmann est un réseau de neurones stochastique ayant des connexions symétriques entre ses différentes unités (Hinton, et al., 1983). La version originale de cette machine contient seulement des unités visibles. Or, l'introduction des variables cachées dans ce modèle a largement augmenté sa capacité de modélisation des données les plus complexes, et ce, même si elles ne sont pas entièrement visibles.

Afin de faciliter l'entrainement de ce genre de modèles, certaines restrictions sont imposées au modèle original des Machines de Boltzmann. La version la plus répandue est celle dénommée la Machine de Boltzmann restreinte (RBM) (Smolensky, 1986). Les RBM sont caractérisées principalement par l'existence de deux couches de neurones, la première ne

contient que des unités visibles tandis que la deuxième ne contient que des unités cachées. De plus, les RBM se caractérisent par l'absence totale de toute connexion entre les unités d'un même niveau (c.-à-d. connexion entre les neurones visibles-visibles ou bien cachés-cachés). Ces restrictions ont principalement l'avantage de rendre l'apprentissage des RBM assez facile et robuste. Par ailleurs, les RBM constituent l'unité de base des modèles très complexes, tels que les Machines de Boltzmann profondes (DBM) et les Réseaux de croyance profonds (DBN). Dans le cadre de cette thèse, nous ne fournirons pas les détails mathématiques et les algorithmes d'apprentissage de ces modèles. Le lecteur intéressé par le cadre général des Machines de Boltzmann peut se référer à (Hinton, et al., 2006)(Bengio, et al., 2007)(Salakhutdinov, 2009). De plus, si le lecteur est intéressé par l'application de ces modèles au problème de la vérification du locuteur, il est fortement recommandé de se référer à (Stafylakis, et al., 2012a)(Senoussaoui, et al., 2012)(Stafylakis, et al., 2012b)(Vasilakakis, et al., 2013).

Au cours des dernières années, les Machines de Boltzmann ont gagné en popularité dans le domaine du traitement automatique de la parole (Dahl, et al., 2010)(Mohamed, et al., 2011)(Hinton, et al., 2012)(Deng, et al., 2013). L'introduction de ces modèles dans le domaine de la vérification du locuteur est marquée principalement par la publication de ces deux communications (Stafylakis, et al., 2012a)(Senoussaoui, et al., 2012), ainsi que par d'autres travaux publiés ultérieurement (Stafylakis, et al., 2012b)(Vasilakakis, et al., 2013). Dans ces travaux, plusieurs variantes de Machines de Boltzmann ont été appliquées dans l'espace des i-vecteurs afin de traiter le problème de la vérification du locuteur. Les performances obtenues par l'application de ces machines dans l'espace des i-vecteurs n'ont malheureusement pas dépassé celles de l'état de l'art actuel (c.-à-d. les résultats obtenus via l'Analyse discriminante linéaire probabiliste PLDA). L'utilisation des i-vecteurs comme entrées à ces machines ne serait probablement pas le bon choix, du fait que ces machines cherchent à modéliser les corrélations inter-variables, cependant, les i-vecteurs sont connus par leurs variables décorrélées.

1.2.4 Décision

Comme la plupart des systèmes intelligents, un système de reconnaissance du locuteur est contraint de prendre une décision. Dans le cas de la vérification du locuteur, la décision est une *acceptation* lorsque le système conclut l'authenticité de l'identité proclamée ou bien un *rejet* dans le cas contraire. En règle générale, la décision est prise dans la pratique en comparant un score à un seuil de décision (voir la Figure 1.3). Le score est un scalaire produit par le système pour mesurer le degré de ressemblance entre deux segments vocaux (enrôlement/test). Plus le score est élevé, plus l'incertitude concernant l'appartenance des deux segments au même locuteur est faible et vice-versa.

1.2.5 Normalisation des scores

En observant le processus de vérification, nous constatons que ce dernier est composé d'une suite d'opérations de réduction de dimensionnalité (c.-à-d. de supervecteur de grande dimension jusqu'au score scalaire). En fait, chaque étape réduit principalement les dimensions qui représentent la variabilité indésirable. En dépit de la multitude des étapes filtrantes, la variabilité indésirable peut se propager jusqu'aux scores (voir la Figure 1.3). En conséquence, plusieurs méthodes ont été proposées dans la littérature pour compenser ces variabilités. L'hypothèse principale de toutes les méthodes de normalisation stipule que la distribution des scores des clients et celle des scores des imposteurs suivent des lois normales unidimensionnelles. Dans le but d'atténuer la variabilité au niveau des scores, les méthodes de normalisation visent à standardiser la distribution des imposteurs de la manière suivante :

$$score^{(norm)} = \frac{score - \mu_{imp}}{\sigma_{imp}} \tag{1.20}$$

où $score^{(norm)}$ est le score obtenu par la normalisation du score brut en utilisant la moyenne μ_{imp} et la déviation standard σ_{imp} estimées à partir des scores d'une cohorte des modèles d'imposteurs.

1.2.5.1 z-norm

Le but de la méthode *z-norm* (Li *et al.* 1988) est d'atténuer les biais causés par les différentes qualités d'estimations des modèles de locuteurs. En effet, la qualité d'estimation d'un modèle du locuteur est essentiellement liée à la longueur, au nombre et à la qualité audio des segments d'enrôlement de ce locuteur. Les paramètres de la normalisation dans ce cas sont estimés en dehors de l'étape du test (c.-à-d. un calcul hors-ligne) et ce, en calculant les scores d'un ensemble d'énoncés des imposteurs vis-à-vis le modèle du locuteur cible.

1.2.5.2 t-norm

La normalisation *t-norm*, quant à elle, vise à réduire la variabilité intersessions propagée jusqu'aux scores. Le mécanisme derrière cette méthode consiste à estimer les paramètres de la normalisation (c.-à-d. la moyenne et la déviation standard) en calculant le score du segment du test vis-à-vis un ensemble de modèles des imposteurs.

1.2.5.3 s-norm

Un des aboutissements dans le domaine de la vérification du locuteur après l'introduction d'espace d'i-vecteur est de rendre symétrique le calcul du score entre un modèle du locuteur cible et un segment de test. De ce fait, une méthode tenant en compte cet aspect est proposée dans la littérature sous l'appellation *s-norm* (Kenny, 2010a). La formule de la normalisation symétrique est donnée comme suit :

$$score^{(norm)} = \frac{score - \mu_e}{\sigma_e} + \frac{score - \mu_t}{\sigma_t} \tag{1.21}$$

où μ_e, σ_e, μ_t et σ_t sont les moyennes et les déviations standards estimées à partir d'un ensemble de scores des segments des imposteurs calculés vis-à-vis le segment d'enrôlement et celui du test.

1.2.6 Calibration et fusion des scores

La dernière étape du processus de la reconnaissance du locuteur est nommée la *calibration*. Cette dernière consiste principalement à fixer le seuil de décision optimal et unique afin de séparer les scores des imposteurs des scores des clients (voir Figure 1.3). La façon la plus simple de sélectionner ce seuil est d'utiliser un ensemble de développement afin de l'estimer empiriquement tout en minimisant la fonction du coût de détection (*Detection Cost Fonction*, DCF). Ce seuil est utilisé par la suite pour prendre la décision lors du test. Une deuxième stratégie de la calibration consiste à transformer les scores bruts d'un système de vérification en des scores qui se comportent comme de vrais *logarithmes de rapports de vraisemblances* (*Log Likelihood Ratio* , LLR). Suite à cette transformation, la théorie bayésienne peut nous fournir la formule exacte du seuil de décision θ indépendant des vecteurs de caractéristiques comme suit :

$$\begin{aligned} \log\left(\frac{P(x|H_0)}{P(x|H_1)}\right) &\geq \theta \\ &\geq \log\left(\frac{p_0}{1-p_0}\right) + \log\left(\frac{C_{FR}}{C_{FA}}\right) \end{aligned} \tag{1.22}$$

où x est un essai de vérification[5], $P(x|H_0)$ et $P(x|H_1)$ sont respectivement les probabilités de l'hypothèse H_0 (c.-à-d. que les deux segments de x soient émis par un même locuteur) et de l'hypothèse H_1 (c.-à-d. qu'ils sont émis par deux locuteurs différents), P_0 est la probabilité *a priori* d'apparition de H_0 et 1 - P_0 est celle d'apparition de H_1, finalement, C_{FR} et C_{FA} sont respectivement les coûts encourus pour un *faux rejet* et une *fausse acceptation*.

Outre que la normalisation et la calibration, les scores des systèmes de la reconnaissance du locuteur sont aussi sujets à une autre opération nommée la *fusion*. Par le mot *fusion*, nous entendons une façon de combiner les scores $s_i(x)$ d'un essai de vérification x, calculés par N différents systèmes de vérification afin de produire un seul score s_f. Bien qu'il existe diverses

[5] Notez qu'un essai de vérification est composé de deux segments de la parole.

architectures de combinaisons entre les systèmes de reconnaissance de formes en général (Campbell, *et al.*, 2007)(Kajarekar, 2005)(Ferrer, 2006), il est bien répandu d'adopter des combinaisons au niveau des scores dans le cas des systèmes de la vérification du locuteur, notamment la formule linéaire de la fusion (Brummer, *et al.*, 2006) :

$$\begin{aligned} s_f &= s(x, \mathbf{w}) \\ &= w_0 + \sum_{i=1}^{N} w_i s_i(x) \end{aligned} \tag{1.23}$$

où $\mathbf{w} = \{w_i\}_{i=1..N}$ est l'ensemble des poids de la fusion.

Récemment, la régression logistique est devenue la méthode la plus adoptée pour entrainer les poids de la fusion (Pigeon, *et al.*, 2000)(Brummer, *et al.*, 2006)(Brummer, *et al.*, 2007)(Van Leeuwen, *et al.*, 2007). À la différence de la manière habituelle de son utilisation dans divers problèmes de la reconnaissance des formes, la régression logistique est utilisée dans le domaine de la vérification du locuteur pour produire des scores qui se comportent comme des *logarithmes de rapports de vraisemblances* LLR au lieu des probabilités *a posteriori de logit*.

1.2.7 Corpus de données et mesures d'évaluation

Il est bien connu que l'institut américain NIST[1] a énormément influencé, voir même dominé, les recherches dans le domaine de la reconnaissance du locuteur, notamment ces deux dernières décennies, et ce, en fournissant des corpora de données et en organisant des campagnes d'évaluations à l'échèle mondiale. Dans une stratégie d'évaluation dite sévère telle que les campagnes de NIST, le protocole d'évaluation dicte qu'un système de vérification doit subir un nombre relativement élevé (des milliers) de tests. Ces tests sont généralement prescrits sous forme de listes d'essais et chacun de ces essais se constitue d'un premier segment d'enrôlement portant le code de son locuteur client et d'un deuxième segment du test dont le locuteur est inconnu au système. Les systèmes des participants dans les campagnes de NIST doivent produire un score pour chaque essai prescrit dans ces listes et

les envoyer à NIST pour procéder à l'évaluation des performances et au classement des systèmes par la suite.

1.2.7.1 Mesures d'évaluation

Un système de vérification du locuteur est susceptible de produire deux types d'erreurs (voir Figure 1.3) :

- *Fausse acceptation (FA)* : quand il authentifie à tort une identité proclamée.
- *Faux rejet (FR)* : quand il rejette à tort un client authentique.

Toute mesure fiable des performances des systèmes de la vérification du locuteur doit impérativement tenir compte de ces deux types d'erreurs. Étant donné un ensemble de scores produit par un système donné, nous pouvons calculer le taux des fausses acceptations (T_{FA}) et le taux des faux rejets (T_{FR}) pour chaque seuil de décision $\theta \in]-\infty,+\infty[$ comme suit :

$$T_{FA} = \frac{Nombre\ de\ FA}{Nombre\ des\ accès\ imposteurs} \tag{1.24}$$

$$T_{FR} = \frac{Nombre\ de\ FR}{Nombre\ des\ accès\ clients} \tag{1.25}$$

Ces deux taux d'erreurs seront combinés pour avoir les deux mesures d'évaluation les plus répandues dans le domaine de la vérification du locuteur de la façon suivante :

- *Equal Error rate (EER)* : c'est le point correspondant à l'égalité entre les deux taux erreurs

$$T_{FA} = T_{FR} \tag{1.26}$$

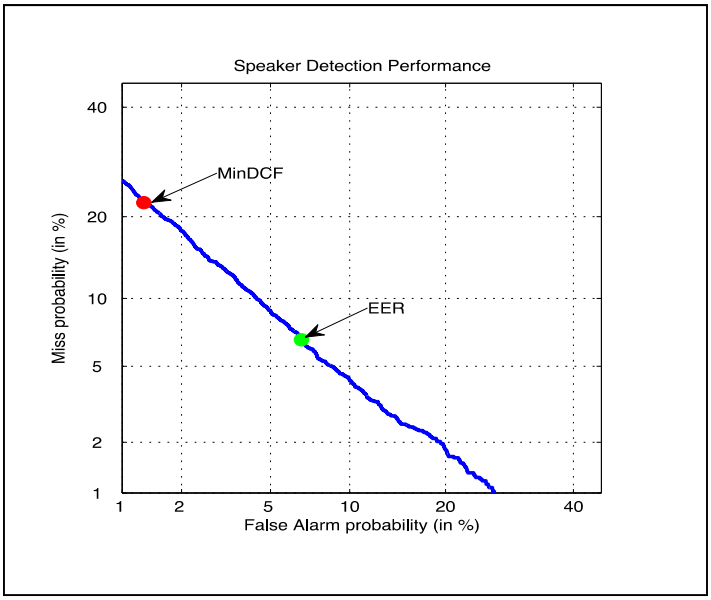

Figure 1.4 Illustration graphique DET de NIST sur laquelle l'ERR et le minimum DCF sont respectivement représentés par le point vert et le point rouge.

– *Minimum of Detection Cost Function (minDCF[6])* : est une fonction des coûts définie par :

$$DCF = C_{FR} T_{FR} P_{client} + C_{FA} T_{FA} (1 - P_{client}) \tag{1.27}$$

où C_{FR} et C_{FA} sont respectivement les coûts relatifs au faux rejet et à la fausse acceptation, le P_{client} est la probabilité *a priori* qu'une identité proclamée soit authentique. Il est à noter que les coûts C_{FR} et C_{FA} et la probabilité *a priori* P_{client} sont fixés par l'utilisateur selon ses besoins, de plus, P_{client} ne reflète pas la probabilité *a priori* calculée sur un ensemble de test ou de développement.

Outre que ces deux mesures, NIST a également proposé une illustration graphique mise à disposition des chercheurs et des développeurs afin de bien analyser les résultats de leurs systèmes de la vérification du locuteur. En fait, cette courbe est connue sous le nom *DET* (*Detection Error Tradeoff*). La courbe *DET* est obtenue par l'évaluation des T_{FR} et T_{FA} en

[6] La version normalisée de la fonction DCF est celle considérée par NIST[1].

faisant varier le seuil de décision sur l'ensemble des scores, ainsi, cette courbe fournit une vue d'ensemble des performances du système.

Tableau 1.1 Exemple des tâches telles que prescrites pour NIST-SRE 2010

		Test		
		10sec	1conv	1conv *sum*[7]
Enrôlement	10sec	Optionnelle		
	1conv	Optionnelle	**Obligatoire**	Optionnelle
	8conv		Optionnelle	Optionnelle
	8conv *sum*		Optionnelle	Optionnelle

1.2.7.2 Tâches des évaluations NIST

Durant chaque campagne d'évaluation de NIST, un protocole d'évaluation englobe un ensemble de tâches qui sont déterminées en fonction de la durée des segments d'enrôlement et celle des segments du test (voir Tableau 1.1). Généralement, il y a seulement une tâche à accomplir obligatoirement par tous les participants (voir la case grise du Tableau 1.1).

Selon le type du canal et la façon dont les enregistrements sont produits, NIST distingue deux catégories de fichiers audio :

1) *Fichiers téléphoniques* : dans cette catégorie on distingue deux types de fichiers :
    a. ***Tel*** : une conversation téléphonique simple contenant les deux côtés de la conversation, nommée aussi en anglais « *4-wires file* ».
    b. ***Summed*** : des fichiers audio obtenus par l'adition, échantillon par échantillon, des deux cotés d'une conversation « *4-wires* », sont aussi nommés « *2-wires files* ».

[7] Le mot *sum* est pour indiquer un type de fichiers audio obtenus par la sommation échantillon par échantillon des deux cotés « *channel sides* » d'un appel téléphonique. Le mot *conv* indique une conversation d'une durée approximative de deux minutes et demie.

2) *Fichiers microphoniques* : nous distinguons aussi deux types de fichiers :
    a. ***Mic*** : des conversations téléphoniques enregistrées par des microphones.
    b. ***Interview*** : des enregistrements contenant la parole d'une personne qui constitue le locuteur cible et la parole d'un intervieweur enregistrés par un microphone posé sur la table d'entrevue. L'intervieweur porte un microphone-proche porté sur sa tête.

Tableau 1.2 Exemple des cinq premières conditions « det » de NIST-SRE 2010

	Canal d'enrôlement	*Canal du test*	*Restriction*
det1	Interview	Interview	Mêmes microphones
det2	Interview	Interview	Microphones différents
det3	Interview	Tel	-
det4	Interview	Mic	-
det5	Tel	Tel	Téléphones différents

En fonction des types du canal utilisé pour l'enrôlement et le test, la tâche principale de la campagne de NIST-SRE 2010 se divise en plusieurs conditions dont chacune porte le nom « *det* » en plus d'un numéro, par exemple « *det1* » (voir les 5 premières conditions dans le Tableau 1.2).

1.2.7.3 Données d'entrainement de NIST

Au fil des années, NIST a accumulé une quantité importante de données d'entrainement (voir Tableau 1.3). En effet, ces données sont à la disposition des participants dans les évaluations de NIST dans le but de les utiliser principalement à des fins d'estimation des paramètres des classificateurs (GMM, JFA, PLDA, etc.) et des extracteurs des vecteurs de caractéristiques (i-vecteurs).

Tableau 1.3 Durées en heures des données d'entrainement de NIST.

	Nom du corpus	**Durée (heures)**
Données microphoniques	MIX05	223
	MIX06	230
Données téléphoniques	Fisher	3916
	MIX04	396
	MIX05	200
	MIX06	815
	Switchboard II	1186
	Switchboard cellulaire	662

1.3 Regroupement en locuteurs dans les grandes bases de données (Clustering)

1.3.1 Définition et utilité de la tâche

Le regroupement ou ce qu'on appelle souvent la classification automatique (*clustering*) est un problème traditionnellement connu dans plusieurs domaines, entre autres, dans la reconnaissance des formes et de l'apprentissage-machine. Dans un ensemble de données non étiquetées, l'objectif d'une tâche de regroupement est de relier les observations les plus proches en terme d'une métrique adoptée, et ce, afin de déterminer la structure intrinsèque de ces données. Cette tâche devient plus complexe lorsqu'on ignore le nombre et la forme des distributions des classes de l'ensemble de données à regrouper.

Dans le cas du traitement de la parole, le regroupement en locuteurs d'un ensemble de segments audio non étiquetés consiste à attribuer à chaque segment un identifiant correspondant à son locuteur émetteur. Il est à souligner que chaque segment audio est présumé contenir la parole d'un unique locuteur. Le regroupement en locuteurs peut être considéré comme un but en soi quand il s'agit par exemple du regroupement d'un ensemble des enregistrements dont chacun contient la parole d'un seul locuteur. Il est aussi considéré comme une sous-tâche dans le cas de la structuration en tours de parole d'un seul flux audio multilocuteur par exemple, et dans ce cas-là, une étape préalable de la segmentation de ce flux où chaque segment contient la parole d'un seul locuteur est indispensable. Les effets du

canal entre les segments à regrouper constituent la principale différence entre ces deux façons d'exploitation du regroupement en locuteurs. Dans le premier cas, chaque segment est considéré comme un enregistrement indépendant. Ce qui implique que les segments d'un même locuteur sont enregistrés sur différentes sessions, et ainsi, une variabilité indésirable complique la tâche de regroupement. Dans le cas de la structuration en tours de parole, le scénario est considérablement différent, du fait que tous les segments sont enregistrés lors d'une même session. Ainsi, les effets du canal dépendant du locuteur (le type du microphone ou du téléphone, le bruit du fond, le positionnement du locuteur par rapport au microphone, etc.) peuvent jouer un rôle positif dans la procédure de la distinction entre les locuteurs participants.

Le regroupement en locuteurs, qu'il soit pour les grandes corpora de données ou pour un seul flux (c.-à-d. la structuration en tours de parole), est une discipline substantielle du traitement de la parole. Il fournit une solution adéquate pour l'extraction des métadonnées afin d'étiqueter automatiquement un corpus de données. Ces données peuvent servir à l'adaptation non supervisée des modèles indépendants du locuteur dans le but d'améliorer leurs performances. Cette discipline peut ainsi servir à faciliter l'archivage, le stockage et la recherche dans les grandes bases de données audio.

1.3.2 Corpora de données et mesures d'évaluation

Le regroupement en locuteurs vise à assigner chaque segment à une classe (groupe) de façon que : 1) chaque groupe contient les segments d'un même locuteur, 2) tous les segments d'un même locuteur assignés au même groupe (voir Figure 1.5). En se basant sur ces deux concepts, deux mesures de performances d'un regroupement sont définies dans la littérature (Van Leeuwen, 2010), à savoir, l'impureté de classe (I_c) (*Cluster Impurit*y) et l'impureté du locuteur (I_s) (*Speaker Impurity*). Il est à noter que le mot « classe » dans I_c fait référence au groupe des segments assemblés par l'algorithme du regroupement et que le mot « locuteur » dans I_s fait référence à la vraie identité du locuteur émetteur d'un ensemble de segments.

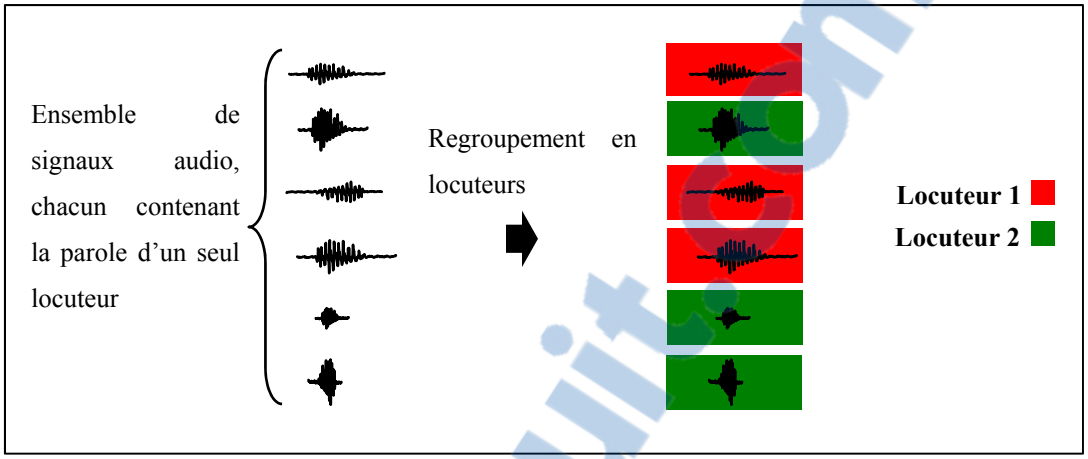

Figure 1.5 Regroupement en locuteurs d'un ensemble d'enregistrements audio.

Afin de pouvoir fournir les formules mathématiques de ces deux mesures, nous devons définir les concepts suivants :

- $\{h_i\}_{i=1..H}$: l'ensemble des classes hypothétiques fournies par l'algorithme de regroupement en locuteurs.
- $R(j)$: le vrai locuteur (référence) du segment j.
- $C(j)$: la classe à laquelle le segment j était assigné.
- S_k : l'ensemble des segments émis par le locuteur k.
- $f_{ik} = f_k(R(h_i))$: la fréquence d'occurrence des segments du locuteur référence k dans la classe h_i, ces fréquences sont triées en ordre décroissant dans la classe h_i.
- $n_i = \sum_k f_{ik}$: est le nombre des segments de la classe h_i.
- $N = \sum_i n_i$. : est le nombre total des segments à regrouper.
- $g_{ki} = g_i(C(S_k))$: la fréquence d'apparition des segments du locuteur k dans la même classe h_i, ces fréquences sont ainsi triées en ordre décroissant pour chaque locuteur k.
- $m_k = \sum_i g_{ki}$: est le nombre des segments de la classe h_i.
- $N = \sum_k m_k$: est le nombre total des segments à regrouper.

Les mesures d'impureté sont alors définies comme suit :

– ***Impureté de classe*** :

$$I_c = 1 - \frac{1}{N}\sum_i f_{i1} \tag{1.28}$$

– ***Impureté du locuteur*** :

$$I_s = 1 - \frac{1}{N}\sum_k g_{k1} \tag{1.29}$$

Il faut noter que si chaque classe contient un seul segment (c.-à-d. le nombre des classes égales au nombre des segments) nous obtenons une impureté de classe nulle. En revanche, l'affectation de tous les segments à une seule classe engendre une impureté du locuteur nulle. Nous observons l'existence d'une relation inverse qui relie le couple d'impuretés (I_c, I_s), ainsi, les performances optimales d'un système de regroupement sont exprimées en fonction d'un point de compromis entre les deux impuretés (i.e. $I_c = I_s$). Une telle relation peut être considérée comme similaire à celle qui relie les deux types d'erreurs du système de la vérification du locuteur (*FA*, *FR*), donc, nous pouvons aussi adopter la courbe *DET* afin de mieux observer l'évolution du couple (I_c, I_s).

1.3.2.1 Corpus de données

Au cours de ce travail, l'intérêt accordé à la tâche du regroupement en locuteurs est plus particulièrement un intérêt de validation dans le sens de vouloir tester l'efficacité d'un algorithme du regroupement en locuteurs. De ce fait, nous avons adopté le corpus des données téléphoniques fournies par NIST dans le cadre de sa campagne d'évaluation de la reconnaissance du locuteur en 2008.

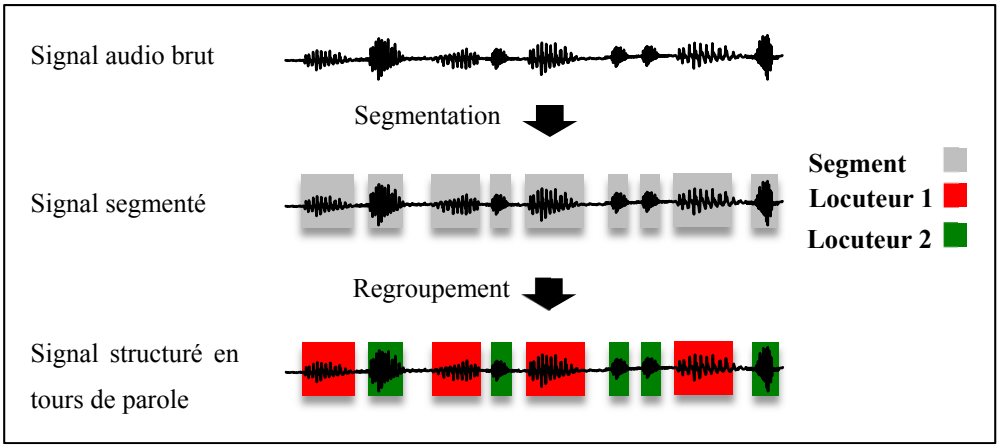

Figure 1.6 Structuration en locuteurs d'un flux audio contenant deux locuteurs.

1.4 Structuration en tours de parole (Diarization)

1.4.1 Définition et utilité de la tâche

La structuration en tours de parole (STP) d'un flux audio multilocuteurs est la tâche de fractionner ce flux en des segments contenants la parole d'un seul locuteur et de les regrouper par la suite en locuteurs telle que présentée dans la section précédente. Ainsi, des réponses sont possibles aux questions « Qui parle ? Quand ? » pour ce flux audio (voir l'exemple d'un fichier audio de deux locuteurs sur la Figure 1.6).

Outre que les domaines d'application du regroupement en locuteurs cités précédemment (voir Section 1.3.1) et qui sont communs avec la structuration en tours de parole, ce dernier peut servir également à l'automatisation de plusieurs tâches qui s'avèrent indispensables de nos jours. Nous citons à titre d'exemple la transcription automatique des enregistrements audiovisuels tels que les cours en ligne[8], les conférences, les réunions, etc. Cette transcription peut être suivie par une traduction vers plusieurs langues différentes de la langue de départ de l'enregistrement. De plus, dans certains enregistrements multilocuteurs, la vérification d'un

[8] Un exemple des cours en ligne se trouve : http://ocw.mit.edu/courses

locuteur parmi ceux enregistrés n'est possible qu'après l'exécution de la structuration de ce fichier en tours de parole afin de séparer la parole de chacun de ces locuteurs.

La difficulté de la tâche de la structuration en tours de parole s'accentue avec l'absence de toute information *a priori* concernant le nombre de locuteurs impliqués dans le fichier à segmenter. Habituellement, cette information n'est pas fournie dans le cas des enregistrements broadcastés, des réunions et des conférences. Cependant, elle pourrait être disponible dans le cas des conversations téléphoniques (Kenny, *et al.*, 2010b).

Comme il est mentionné à maintes reprises précédemment dans ce chapitre, la structuration en tours de parole se divise en deux sous-tâches principales, à savoir, la segmentation et le regroupement. Dans ce qui suit, nous fournirons plus de détails sur ces deux étapes tout en faisant référence à la méthodologie répondue pour accomplir chacune de ces tâches.

1.4.1.1 Segmentation

Comme son nom l'indique, cette étape vise à fragmenter le signal audio brut en des zones, souvent appelées segments ou tours de parole (*speaker turn*), qui sont censées contenir la parole d'un unique locuteur. Pour ce faire, la segmentation cherche à détecter les points du changement. L'idée de base des approches traditionnelles de la segmentation est de fixer successivement deux fenêtres glissantes, avec un éventuel chevauchement, sur les trames acoustiques. Ainsi le point du chevauchement entre les deux segments (fenêtres) est considéré comme un point de changement du locuteur lorsque la distribution des trames bornées par la première fenêtre est différente de celle des trames bornées par la deuxième fenêtre. En revanche, si les deux distributions sont jugées similaires, en se basant dans les deux cas sur une métrique et un seuil de décision, les deux segments sont considérés comme étant générés par le même modèle (locuteur). On glisse alors les deux fenêtres pour refaire le même test afin de parcourir le reste du signal.

La méthode de segmentation la plus répandue dans la littérature est celle connue sous l'acronyme BIC (*Bayesien Information Criterion*) (Schwarz, 1978)(Chen, *et al.*, 1998). BIC

est un critère de sélection bayésien du modèle en se basant sur la maximisation de sa vraisemblance pénalisée par sa complexité. La complexité des modèles est souvent mesurée par le nombre de ses paramètres libres. Dans le cas du concept de la segmentation décrit dans le paragraphe précédent, le critère BIC est utilisé dans ce scénario pour sélectionner laquelle des deux modélisations est la plus adéquate pour les deux segments consécutifs : la modélisation par deux modèles distincts ce qui implique l'existence d'un point du changement ou bien la modélisation par un seul modèle ce qui implique l'absence du changement du locuteur.

Une segmentation plus élémentaire est habituellement considérée comme une étape préalable à celle décrite précédemment. Cette segmentation est basée sur les bornes « silence/parole » fournies par un détecteur de silence. Certainement, un segment limité par deux silences successifs ne contient pas forcément la parole d'un unique locuteur. Ainsi une telle segmentation nécessite l'exécution d'une méthode de segmentation beaucoup plus rigoureuse. Cependant, dans le cas de la parole téléphonique, une segmentation uniforme de courte durée, approximativement une seconde, est utilisée avec succès comme une segmentation initiale dans plusieurs travaux (Kenny, *et al.*, 2010b)(Shum, *et al.*, 2013)(Senoussaoui *et al.*, 2014).

1.4.1.2 Regroupement

Hormis la façon avec laquelle on traite la variabilité intersessions entre les segments à regrouper, cette étape de regroupement est similaire à celle du regroupement en locuteurs dans les grands corpora de données (voir la Section 1.3). Contrairement au regroupement en locuteurs, la variabilité intersessions est avantageuse pour la tâche de la distinction entre les locuteurs dans le cadre de la structuration en tours de parole d'un flux audio. Ceci est dû au fait que tous les segments d'un même locuteur sont enregistrés durant la même session. Ainsi, la variabilité dépendante de chaque locuteur (c.-à-d. la variabilité due au type de microphone, à l'environnement d'enregistrement, à la qualité de transmission, etc.) peut contribuer à l'augmentation de la variabilité interlocuteurs.

Les méthodes de la classification ascendante/descendante étaient les mécanismes les plus utilisés pour accomplir la tâche de la classification automatique (c.-à-d. *clustering*). Le principe de la classification hiérarchique est facile, il consiste à regrouper/séparer deux observations après chaque itération de l'algorithme. Le processus de regroupement/séparation est basé sur une mesure de ressemblance/dissemblance. Cette mesure pourrait être une distance, une vraisemblance ou bien même un score d'un système de vérification du locuteur. Récemment, des méthodes bayésiennes variationnelles sont ainsi émergées dans ce domaine (Valente, 2005). L'idée de base de ces méthodes bayésiennes consiste toujours à la sélection du modèle le plus adéquat étant donné un ensemble d'observations. Dans la pratique, le modèle de mélange de gaussiennes GMM était adopté afin de modéliser le mélange des locuteurs d'un flux audio dans l'espace des i-vecteurs (Shum, *et al.*, 2013). Ainsi, la parole de chaque locuteur dans le flux est modélisée par une gaussienne de ce mélange. En effet, le nombre de locuteurs est *a priori* inconnu, alors, une limite supérieure $L^{(\mathrm{sup})}$ de ce dernier est fixée. Pour chaque nombre (entier) de locuteurs $1 \le L \le L^{(\mathrm{sup})}$, un modèle de L gaussiennes sera entrainé. Dans le but de détecter le nombre réel de locuteurs, l'énergie libre (*free energy*) est utilisée comme critère de sélection du modèle le plus adéquat aux données. Ainsi, le nombre de ses composantes reflète le nombre de locuteurs. Cette architecture a aussi été proposée avec un plus grand degré de complexité afin de fonctionner dans l'espace des vecteurs acoustiques (Valente, 2005). Dans cet espace, chaque locuteur est modélisé par un GMM au lieu d'une simple gaussienne. Quant au modèle du mélange, il est remplacé par un modèle de Markov caché (HMM) ergodique.

Un peu plus tard, la modélisation des locuteurs par l'analyse en facteurs FA (*Factor Analysis*) a remplacé le GMM dans le mécanisme de la sélection bayésienne du modèle afin de donner naissance à un nouveau modèle qui a particulièrement marqué son succès dans la structuration en tours de parole d'un flux audio à deux locuteurs (Kenny, *et al.*, 2010b).

Outre que les méthodes hiérarchiques et bayésiennes, une méthode non paramétrique nommée l'algorithme de décalage de la moyenne (*Mean Shift*) a été aussi proposée (Stafylakis, *et al.*, 2010)(Stafylakis, *et al.*, 2012). Dans cette thèse, nous proposons une nouvelle version de cet algorithme testé avec succès sur la structuration en tours de parole

téléphonique. Nous fournirons les détails de l'algorithme original ainsi que ceux de la nouvelle version avec ses deux mécanismes de regroupement dans le Chapitre 6.

Dans la littérature, on trouve de nombreuses autres méthodes de regroupement. Pour les lecteurs qui s'y intéressent, de très bonnes revues peuvent être trouvées dans (Tranter, *et al.*, 2006)(Kotti, *et al.*, 2008)(Moattar, *et al.*, 2012).

1.4.2 Évaluation des performances

Les performances d'un système de structuration en tours de parole peuvent être évaluées avec les mesures d'impureté présentées dans la section 1.3.2. Cependant, une mesure connue sous l'acronyme DER (*Diarization Error Rate*) proposée par NIST est devenue la mesure la plus utilisée dans ce domaine. D'ailleurs NIST a également fourni un script[9] pour évaluer facilement le DER d'une façon unique et standard. En fait, le DER est obtenu par l'addition de trois types d'erreurs commises par un système de la structuration en tours de parole, ces erreurs sont les suivantes :

- *ERR-S* : confusion de la parole d'un locuteur avec la parole des autres.
- *FA* : une fausse acceptation d'un segment de silence.
- *MISS* : un rejet d'un segment de la parole.

Le nombre des locuteurs détectés (*Number of Detected Speakers*, NDS) est un autre indice de performance très important, qui n'est malheureusement pas largement adopté par la communauté de recherche. Le nombre NDS est associé à chaque erreur calculée (DER), il donne une information supplémentaire pour évaluer le comportement du système. Un système performant est celui qui optimise simultanément le DER et le NDS.

Pratiquement tous les systèmes de structuration en tours de parole et du regroupement se basent au moins sur un hyper-paramètre contrôlant le nombre de classes (c.-à-d. le nombre de locuteurs) détectées NDS. Certainement, le changement de cet hyper-paramètre affecte

[9] http://www.nist.gov/speech/tests/rt/rt2006/spring/code/md-eval-v21.pl

également le DER, ainsi, un graphe visualisant la variation de DER par rapport au NDS est aussi un outil efficace pour évaluer les performances de ces systèmes.

1.4.2.1 Corpus de données

Dans le cadre de cette thèse, nous nous intéressons plus particulièrement au problème de la structuration en tours de parole des flux audio téléphoniques. L'une des caractéristiques du discours téléphonique est la durée relativement courte des tours de parole (approximativement une seconde) (Kenny, *et al.*, 2010b). Une des conséquences directes de la brièveté des tours de parole est la difficulté de représenter avec précision ces segments dans l'espace des i-vecteurs. Nous étudierons dans le cadre de ce travail des méthodes simples nous permettant de surmonter cette difficulté tout en obtenant des résultats comparables à ceux obtenus par des méthodes de l'état de l'art dans ce domaine (Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Vaquero, 2011)(Shum, *et al.*, 2012).

Le corpus de données *CallHome* sera notre base de validation pour la tâche de la structuration en tours de parole. Ce corpus de données a été fournie par NIST pour accomplir la tâche de la segmentation durant la campagne d'évaluation de la reconnaissance du locuteur (Martin, et al., 2001). *CallHome* est un corpus multilingue (six langues) de la parole téléphonique. Il est séparé en deux sous-ensembles disjoints: un pour le développement contenant 42 conversations de 2 à 4 locuteurs participants par conversation, et un deuxième ensemble, dédié au test, contenant 500 conversations de 2 à 7 locuteurs par conversation.

CHAPITRE 2

REPRÉSENTATION DU SIGNAL VOCAL PAR LES I-VECTEURS

2.1 De la représentation à court terme à la représentation par les i-vecteurs

Dans ce chapitre nous nous focalisons sur la représentation de la voix, une étape très importante pour le processus de la reconnaissance du locuteur. En effet, durant ces quatre dernières années, la représentation de la voix par des i-vecteurs est devenue dominante dans le domaine de la reconnaissance du locuteur (Dehak, *et al.*, 2011b)(Kenny, 2012). Ce chapitre détaille le processus d'extraction des i-vecteurs à partir du signal vocal.

Comme montré au Chapitre 1, le signal de la parole, de par sa nature pseudo-aléatoire et très variable, est souvent incompatible avec la plupart des méthodes standards de la reconnaissance des formes. Par conséquent, ce signal nécessite un prétraitement appelé l'étape du *paramétrage* ou bien l'extraction des vecteurs de caractéristiques. L'hypothèse principale derrière cette étape est que le signal de la parole est quasi stationnaire sur une très courte période (allant typiquement de 10 ms à 30 ms) communément nommée fenêtre du traitement. Ainsi, une image de l'état du signal durant cette période peut être capturée et représentée par un vecteur numérique. En effet, ce traitement local est répété périodiquement (typiquement chaque 10 *ms*) afin d'obtenir une succession de vecteurs dénommés *vecteurs acoustiques*.

Outre que le paramétrage à court terme du signal de parole décrit dans le paragraphe précédent, il existe deux autres classes de paramètres, les paramètres prosodiques qui modélisent principalement le style et le rythme de la parole, et les paramètres à haut niveau qui modélisent l'aspect socioculturel du locuteur (c.-à-d. son environnement, les mots qu'il utilise fréquemment, son dialecte, etc.) (Doddington, 2001).

Le paramétrage du signal constitue alors le premier maillon de la chaine d'extraction des i-vecteurs comme nous le verrons dans la suite de ce chapitre.

2.1.1 Extraction des vecteurs MFCC

Bien qu'il existe plusieurs méthodes d'extraction de vecteurs acoustiques (voir Section 1.2.2), la représentation basée sur une échelle non linéaire nommée l'échelle de Mel (c.-à-d. MFCC) (Davis *et al.*, 1980) est la plus utilisée. En outre, la principale caractéristique de cette échelle est sa simulation du mécanisme perceptuel non linéaire de l'oreille humaine.

La représentation MFCC sera celle utilisée pour la validation de toutes les méthodes proposées dans cette thèse. Dans ce qui suit, nous décrirons brièvement les étapes principales du processus d'extraction (voir Figure 2.1) des vecteurs acoustiques de type MFCC :

- ***Fenêtrage*** : consiste en premier temps à découper le signal en trames chevauchées de faible durée où le signal est considéré comme quasi stationnaire. Ensuite, chaque trame est multipliée par une fenêtre temporelle d'analyse qui peut prendre plusieurs formes, uniformes, triangulaires, gaussiennes, etc. Toutefois, la fenêtre de Hamming reste la plus utilisée dans le domaine du traitement de la parole. En fait, l'objectif principal de l'étape du fenêtrage du signal est d'atténuer les discontinuités du signal au bout de ces trames tout en réduisant le signal à zéro autour de ces extrémités. Le choix de la taille de la fenêtre d'analyse représente toujours un dilemme du fait qu'une fenêtre de très courte durée assure l'hypothèse de la quasi-stationnarité du signal. Cependant elle ne contient pas suffisamment d'échantillons pour assurer une bonne estimation des paramètres du signal.
- ***Transformée de Fourier rapide*** **(Fast Fourier Transform, FFT)** : est un algorithme conçu pour calculer rapidement la transformée de Fourier discrète. La FFT sera appliquée pour chaque fenêtre d'analyse (c.-à-d. trame fenêtrée) afin de réaliser le passage du signal du domaine temporel au domaine fréquentiel.

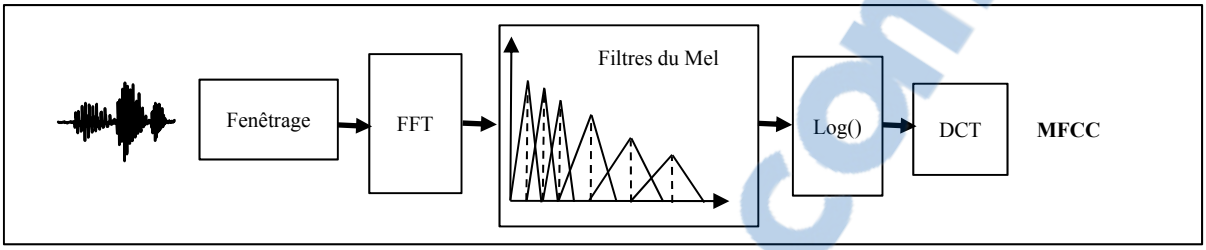

Figure 2.1 Procédure d'extraction des vecteurs caractéristiques à court terme MFCC.

- ***Filtrage selon l'échelle de Mel*** : l'échelle de la perception fréquentielle de l'oreille humaine n'est pas linéaire. Par conséquent, une filtration du signal vocal par une banque de filtres positionnés selon une échelle similaire à la nôtre peut alléger largement la complexité du traitement. L'échelle de Mel est construite à partir d'une série de filtres passe-bandes de formes triangulaires (voir Figure 2.1) positionnés d'une façon linéaire, pour les basses fréquences (< 1000 Hz) et logarithmique pour les hautes fréquences.
- ***Transformée en cosinus discrète*** : cette transformée consiste à multiplier les logarithmes des réponses en énergie des filtres de Mel par des fonctions sinusoïdes de différentes fréquences. L'objectif est de décorréler ces valeurs d'énergie pour constituer par la suite les coefficients (éléments) de notre vecteur MFCC final.

La configuration MFCC adoptée pour la réalisation de toutes les expériences menées dans le cadre de cette thèse est illustrée par le Tableau 2.1.

Tableau 2.1 Configuration d'extraction des vecteurs MFCC adoptée dans les travaux de cette thèse.

Taille de trame	25 *ms*
Pas d'extraction	10 *ms*
Nombre de filtres de Mel	24
F : Dimension de trame	60

À la fin du processus d'extraction des MFCC, nous obtenons une suite de T trames (vecteurs acoustiques) $\mathbf{A}=\{\mathbf{a}_t\}_{t=1..\ T}$. Dans le cas des systèmes de la vérification du locuteur, chaque vecteur acoustique $\mathbf{a}_t$ (de dimension 20 dans notre configuration) est augmenté par une

approximation de ses dérivées temporelles de premier et de second ordres (Δ et ΔΔ). Nous obtenons, ipso facto, des vecteurs de dimension 60. Afin d'atténuer le bruit convolutif du canal de transmission, ces vecteurs acoustiques doivent subir des transformations à base de deux méthodes de compensations très répandues. La première est la normalisation via la soustraction du vecteur moyen de l'ensemble des vecteurs acoustiques (CMS pour Cepstral Mean Subtraction) (Furui, 1981). La deuxième, quant à elle, est une normalisation de type « Feature Warping ». Son objectif est de transformer la distribution des vecteurs acoustiques en une distribution gaussienne sur un intervalle de temps évolutif, typiquement de trois secondes (Pelecanos *et al.*, 2001). La normalisation « Feature Warping » compense le bruit additif et les effets linéaires du canal.

Il est à noter que dans le cas des systèmes de la structuration en tours de parole, seuls les 20 coefficients (19 coefficients de MFCC et un coefficient d'énergie) bruts (c.-à-d. sans aucune normalisation) des vecteurs acoustiques sont utilisés (Reynolds, *et al,*. 2000). Ceci est probablement dû au fait que la plupart des méthodes de normalisation utilisent des fenêtres d'analyse de tailles relativement grandes (typiquement trois secondes), ce qui risque de chevaucher les tours de parole de différents locuteurs.

On souligne également qu'une procédure de détection des régions de silence est indispensable afin de les exclure dans les traitements subséquents.

2.1.2 Modèle du monde (UBM)

La deuxième étape du processus d'extraction des i-vecteurs consiste à construire un modèle acoustique *a priori*. La distribution complexe des vecteurs acoustiques nécessite une modélisation multimodale. En fait, la modélisation par les mélanges de gaussiennes (GMM) est devenue le paradigme le plus répandu dans le domaine de la reconnaissance du locuteur. Le modèle du monde (UBM pour Universal Background Model) est le terme utilisé pour référer au GMM de C composantes gaussiennes entrainé à partir d'une quantité importante (typiquement des centaines d'heures) de vecteurs acoustiques de la parole de plusieurs locuteurs. Comme nous l'avons déjà vu à la Section 1.2.3.1, l'estimation des paramètres de

ce GMM, à savoir $\Omega = \{w_c, \mu_c, \Sigma_c\}_{c=1..C}$, est basée sur le critère du maximum de vraisemblance (ML). Dans le cas du modèle du monde, le choix de ce critère d'estimation est naturel en raison de la disponibilité d'une énorme quantité de données d'apprentissage.

De par sa nature, un extracteur des i-vecteurs ne fait aucune distinction entre les principales sources de la variabilité (voir Section 1.1.1) intrinsèque dans le signal vocal. En fait, un enregistrement d'un locuteur donné est considéré comme indépendant des autres enregistrements, même s'ils appartiennent à ce même locuteur. Par conséquent, un vecteur caché (c.-à-d. un i-vecteur) **x** de dimension $D \times 1$ est associé à chaque enregistrement afin de tenir compte de la variabilité inter-enregistrements uniquement. L'hypothèse principale est que la distribution des vecteurs acoustiques extraits d'un enregistrement donné est un mélange de C gaussiennes dont Σ_c est la matrice de covariance de la $c^{\text{ième}}$ composante, son vecteur moyen est donné comme suit :

$$\mu_c = \mathbf{m}_c + \mathbf{T}_c\mathbf{x} \tag{2.1}$$

où, $\mathbf{m}_c$ est le vecteur moyen de dimension F de la composante c du modèle *a priori* (c.-à-d. l'UBM), la matrice $\mathbf{T}_c$ de dimension $F \times D$ est la $c^{\text{ième}}$ partie de la matrice **T** de dimension $CF \times D$ qui représente la matrice de projection d'espace des i-vecteurs. Elle est aussi souvent connue sous le nom de matrice de variabilité totale.

Donc, le logarithme de la vraisemblance d'un enregistrement vocal représenté par une séquence de vecteurs acoustique $\mathbf{A} = \{\mathbf{a}_t\}_{t=1..T}$, dont l'alignement de chaque vecteur est préalablement donné, peut être écrit tout simplement de la manière suivante :

$$\sum_c \left(N_c \ln \frac{1}{(2\pi)^{F/2} \left|\sum_c\right|^{1/2}} - \frac{1}{2} \sum_t (\mathbf{a}_t - \mathbf{T}_c\mathbf{x} - \mathbf{m}_c)' \sum_c^{-1} (\mathbf{a}_t - \mathbf{T}_c\mathbf{x} - \mathbf{m}_c) \right) \tag{2.2}$$

où le scalaire N_c représente le nombre des trames alignées avec la gaussienne c. Il est à noter que la formule du logarithme de vraisemblance (2.2) peut être évaluée en fonction des statistiques de premier et de second ordres calculés pour chaque gaussienne comme suit :

$$\mathbf{F}_c = \sum_t \mathbf{a}_t \tag{2.3}$$

$$\mathbf{S}_c = \sum_t \mathbf{a}_t \mathbf{a}_t' \tag{2.4}$$

2.1.3 Calcul des statistiques générales

Nous allons utiliser le terme de statistique générale pour faire référence aux statistiques exhaustives ou bien ce qui est connu en anglais par le terme « sufficient statistic ». Étant donné que l'alignement des vecteurs acoustiques n'est réellement pas donné, nous utilisons dans la pratique des statistiques d'ordre zéro, un et deux obtenues par un alignement, de types Viterbi ou Baum-Welch, des vecteurs acoustiques avec le modèle du monde UBM comme suit :

$$N_c = \sum_t \gamma_t^{(c)} \tag{2.5}$$

$$\mathbf{F}_c = \sum_t \gamma_t^{(c)} \mathbf{a}_t \tag{2.6}$$

$$\mathbf{S}_c = \sum_t \gamma_t^{(c)} \mathbf{a}_t \mathbf{a}_t' \tag{2.7}$$

où, $\gamma_t^{(c)} = P(\mathbf{c} \mid \mathbf{a}_t, \Omega)$est la probabilité *a posteriori* (voir équation 1.7) que la trame $\mathbf{a}_t$ ait été générée par la $c^{\text{ième}}$ gaussienne du modèle UBM.

Pour des raisons de simplification des calculs et de réductions du temps d'apprentissage d'extracteur des i-vecteurs, il est considéré dans l'article (Kenny, 2012) que la contribution des statistiques de second ordre dans le processus d'extraction est minime. Par conséquent, cette statistique peut être exclue du calcul du logarithme de vraisemblance. En outre, une autre astuce de simplification a été également proposée dans ce même article (Kenny, 2012). Il s'agit du blanchiment (c.-à-d. la décorrélation) des statistiques de premier ordre en utilisant les paramètres du modèle UBM (les vecteurs moyens et les matrices de covariances) de la manière suivante :

$$\mathbf{F}_c \leftarrow \mathbf{L}_c^{-1}(\mathbf{F}_c - N_c\mathbf{m}_c) \tag{2.8}$$

où, $\mathbf{L}_c\mathbf{L}_c' = \sum_c$ est la décomposition de Cholesky de la matrice de covariance de la $c^{\text{ième}}$ composante d'UBM. En faisant cette normalisation, nous pouvons considérer que $\mathbf{m}_c = \mathbf{0}$ et $\sum_c = \mathbf{I}$ (avec **I** est la matrice identité) pour toutes les composantes, et ce, pour tout le calcul subséquent.

2.1.4 Entrainement de l'extracteur des i-vecteurs

Étant donné un ensemble de statistiques d'ordre zéro et un, extraites à partir d'une collection d'enregistrements d'entrainement alignés avec l'UBM, l'apprentissage d'un extracteur des i-vecteurs consiste à estimer seulement la matrice de la variabilité totale **T** de dimension $CF \times D$. Pour ce qui est des vecteurs moyens $\mathbf{m}_c$ et les matrices de covariances $\sum_c$, ils seront copiés directement de l'UBM (Kenny, 2012).

Avant d'entamer les détails de la procédure d'estimation de la matrice **T,** il est indispensable de présenter les formules de calcul des deux premiers moments de la distribution *a posteriori* du vecteur caché **x** :

$$\langle \mathbf{x} \rangle = \mathrm{Cov}(\mathbf{x}, \mathbf{x}) \sum_c \mathbf{T}_c' \sum_c^{-1} (\mathbf{F}_c - N_c\mathbf{m}_c) \tag{2.9}$$

$$\langle \mathbf{x}\mathbf{x}' \rangle = \mathrm{Cov}(\mathbf{x}, \mathbf{x}) \langle \mathbf{x} \rangle \langle \mathbf{x}' \rangle \tag{2.10}$$

$$\mathrm{Cov}(\mathbf{x}, \mathbf{x}) = \left(\mathbf{I} + \sum_c N_c \mathbf{T}_c' \Sigma_c^{-1} \mathbf{T}_c \right)^{-1} \tag{2.11}$$

Il est à noter que le moment de premier ordre, donné par l'espérance mathématique du vecteur caché **x** présenté par l'équation 2.9, est exactement notre i-vecteur estimé à partir d'un enregistrement donné. En fait, ces calculs (2.9, 2.10 et 2.11) sont les plus importants pour le reste des procédures d'estimation et d'extraction.

L'estimation de la matrice de la variabilité totale est basée sur une forme semblable à l'algorithme EM (*Expectation Maximization*) (Dempster, *et al.*, 1977). Cet algorithme s'implémente en deux étapes principales (Kenny *et al.*, 2005) :

1. **« Expectation »** : cette étape consiste à calculer pour chaque enregistrement vocal *r* les espérances mathématiques de premier et de second ordre données respectivement par les équations 2.9 et 2.10 en utilisant l'estimation courante de la matrice **T**.
2. **« Maximization »** : Dans cette étape nous utilisons les espérances calculées lors de l'étape précédente afin de mettre à jour la matrice **T**.

La mise à jour de la matrice **T** dans la deuxième étape de l'algorithme est basée sur le principe du maximum de vraisemblance. Dans ce cas-là, la formule de la mise à jour de la matrice **T** est la suivante :

$$\mathbf{T}_c = \left(\sum_r \mathbf{F}^{(r)}{}_c \left\langle \mathbf{x}'_{(r)} \right\rangle \right) \left(\sum_r N^{(r)}{}_c \left\langle \mathbf{x}_{(r)} \mathbf{x}'_{(r)} \right\rangle \right)^{-1} \tag{2.12}$$

où, $r = 1..R$ est un indice de R enregistrements d'apprentissage.

Outre que la mise à jour selon le critère du maximum de vraisemblance, il est indispensable d'adopter une deuxième forme de mise à jour basée cette fois-ci sur le critère de la divergence minimale MD (*Minimum Divergence*). L'objectif de cette méthode est de modifier la matrice **T**, de sorte qu'on force la distribution des i-vecteurs pour qu'elle suive une distribution gaussienne standard (c.-à-d. centrée réduite) (Kenny, 2012). Les transformations adéquates à ces fins sont :

$$\begin{aligned} \mathbf{T}_c &\leftarrow \mathbf{T}_c \mathbf{L} \\ \mathbf{x}_{(r)} &\leftarrow \mathbf{L}^{-1} \mathbf{x}_{(r)} \end{aligned} \tag{2.13}$$

où **L** est la matrice triangulaire inférieure obtenue par la décomposition de Cholesky de la matrice suivante :

$$\frac{1}{R} \sum_r \left\langle \mathbf{x}_{(r)} \mathbf{x}'_{(r)} \right\rangle \tag{2.14}$$

Dans la pratique, nous faisons appel à cette normalisation de façon sporadique (par exemple, après chaque cinq itérations de mises à jour selon le critère du maximum de vraisemblance ML). L'algorithme MD est principalement caractérisé par l'accélération de la convergence de l'estimation de la matrice **T**.

Les deux méthodes de mise à jour de la matrice **T** décrites ci-dessus garantissent l'augmentation de la vraisemblance totale des données d'apprentissage calculée selon la *Proposition 2* qui est présentée dans l'article (Kenny *et al.*, 2005).

Il existe aussi une autre version de cet algorithme d'apprentissage de l'extracteur des i-vecteurs basée sur les méthodes bayésiennes variationnelles. Cette version est détaillée dans l'article de Kenny (Kenny, 2012).

2.1.5 Extraction des i-vecteurs

En présence d'un enregistrement vocal et d'une matrice de la variabilité totale, l'extraction du i-vecteur correspondant à cet enregistrement se résume dans les trois étapes suivantes :

1. Extraction des vecteurs acoustiques (incluant leurs normalisations) et élimination des régions de silence.
2. Calcul des statistiques de Baum-Welch d'ordre zéro et un (voir les équations 2.5 et 2.6) de toutes les composantes d'UBM en utilisant les vecteurs acoustiques extraits précédemment.
3. Calcul du i-vecteur correspondant selon l'équation 2.9.

2.2 Compensation des variabilités nuisibles

De par sa nature, un i-vecteur contient toute sorte de variabilités intrinsèques dans le signal vocal d'entrée. Ainsi, une opération de compensation de ces variabilités dépendante de l'application reste indispensable. En effet, cette caractéristique des i-vecteurs est plutôt positive du fait que ces mêmes vecteurs de caractéristiques peuvent être adoptés par plusieurs disciplines à condition d'identifier les sources de la variabilité désirée et celles de la

variabilité nuisible. À titre d'exemple, la variabilité interlocuteurs est primordiale pour les applications de la reconnaissance du locuteur mais nuisible aux applications de la reconnaissance de la parole ou de la reconnaissance de la langue.

De plus, la dimensionnalité modérée des i-vecteurs (typiquement dans les centaines) a largement aidé les scientifiques, dans le sens où elle leur a permis d'utiliser des modèles génératifs ainsi que des méthodes traditionnelles sans se soucier de la malédiction de la dimensionnalité (*curse of dimensionality*) (Bellman, 1957, 2003).

Dans le reste de ce chapitre, nous allons fournir un survol des méthodes de compensation actuellement utilisées dans le domaine de la vérification du locuteur ainsi que dans la structuration en tours de parole des fichiers audio. Les détails mathématiques seront plutôt fournis dans les chapitres qui suivent.

2.2.1 Adaptation des i-vecteurs à la vérification du locuteur

Pour les applications de la vérification du locuteur, les i-vecteurs représentant la parole d'un individu doivent subir des transformations permettant de maximiser la variabilité interlocuteurs et de supprimer ou de minimiser tout autre type de variabilité (c.-à-d. ce qu'on nomme souvent les effets du canal).

Dans le chapitre précédent, nous avons brièvement introduit le modèle génératif PLDA (Kenny, 2010) que nous détaillerons dans les chapitres qui suivent. Ce qui nous intéresse davantage dans cette section, c'est simplement de souligner au lecteur le fait que le PLDA modélise séparément les variabilités (utiles/nuisibles) intrinsèques dans les i-vecteurs afin d'éliminer ce qui est nuisible et de se concentrer sur la variabilité utile pour enfin distinguer la parole des différents locuteurs.

Nous avons vu également dans le Chapitre 1 une méthode de vérification du locuteur basée sur la similarité angulaire (c.-à-d. la similarité du cosinus). Dans un cas pareil, il est essentiel de faire appel aux méthodes classiques d'analyse de données afin d'identifier les transformations nécessaires à la minimisation de la variabilité nuisible des i-vecteurs. Dans la

littérature, la combinaison de l'Analyse discriminante linéaire (*Linear Discriminant Analysis*, LDA) et la normalisation avec la matrice intraclasse (*Within Class Covariance Normalization,* WCCN) est devenue la formule standard de compensation des effets indésirables du canal, et ce, spécialement lorsqu'on se base sur la similarité angulaire du cosinus pour la classification (Dehak, *et al.*, 2010)(Dehak, *et al.*, 2011b)(Senoussaoui, *et al.*, 2013a).

Une normalisation des i-vecteurs via LDA est également appropriée dans le cas du classificateur génératif PLDA (Senoussaoui *et al.*, 2011a). Dans ce cas-là, la normalisation LDA est suivie par une projection des i-vecteurs dans la sphère unitaire afin de réduire la variance totale des données et de rendre leur distribution plus proche d'une distribution gaussienne (Garcia-Romero, 2011).

2.2.2 Adaptation des i-vecteurs à la structuration en tours de parole

Contrairement à la vérification du locuteur, dans les applications de la structuration en tours de parole, la variabilité du canal est plutôt considérée comme favorable. Ceci est dû au fait que l'opération de distinction entre la parole des locuteurs se fait à l'intérieur d'un fichier audio unique contenant la parole de tous ces locuteurs. De ce fait, les segments d'un même locuteur sont tous enregistrés pendant la même session et par conséquent, les effets du canal associés à chaque locuteur peuvent aider à le distinguer des autres.

Une normalisation non supervisée via l'analyse en composantes principales (*Principal Component Analysis,* PCA) a prouvé son efficacité, une fois testée sur un problème de la structuration en tours de parole des communications téléphoniques de deux locuteurs (Shum *et al.*, 2011). En plus que le traitement à base de PCA est non supervisé (c.-à-d. ne nécessite aucun étiquetage de données), il est également local du fait que ce traitement s'effectue fichier par fichier. Ainsi, il a l'avantage de ne solliciter aucune donnée d'apprentissage externe.

Il est également possible d'utiliser les méthodes de compensations supervisées, telles que la LDA et la WCCN (voir section 2.2.1), à condition de définir soigneusement la notion d'*observation* ainsi que celle de la *classe* relatives à notre problème. Comme nous l'avons maintes fois mentionné dans le chapitre 1, la structuration en tour de parole d'un flux audio multilocuteurs se divise en deux étapes successives, à savoir, la segmentation et le regroupement des segments résultant (c.-à-d. tours de parole) en groupes homogènes qui représentent les locuteurs impliqués. Il est ainsi évident de considérer chaque groupe (locuteur) comme une *classe* relative à ce problème et les segments (tours de parole) appartenant à ce groupe comme des observations de la classe en question.

CHAPITRE 3

VÉRIFICATION DU LOCUTEUR

Comme nous l'avons vu précédemment, la tâche de la vérification du locuteur en présence de deux segments de la parole consiste à répondre à la question : « appartiennent-ils au même locuteur ? » De même, nous avons vu la procédure d'une représentation robuste des segments vocaux de durées variables par des vecteurs de dimensions fixes, nommés i-vecteurs. Dans ce chapitre, nous nous focalisons davantage sur les méthodes de la classification des i-vecteurs représentant les voix des locuteurs. Ces méthodes sont adoptées dans l'implémentation des systèmes de l'état de l'art actuel dans le domaine de la vérification du locuteur[10].

3.1 Modèle génératif

3.1.1 Modélisation des i-vecteurs

Étant donné un locuteur *l* ayant un ensemble $\mathbf{X} = \{\mathbf{x}_r\}_{r=1..R}$ de *R* enregistrements représentés par des i-vecteurs de dimension *D*. D'une manière semblable au modèle JFA (Kenny, *et al.*, 2008), un i-vecteur $\mathbf{x}_r$ dépendant du locuteur et du canal peut se diviser en deux composantes (i-vecteurs) de la manière suivante :

$$\mathbf{x}_r = \mathbf{S} + \mathbf{C}_r \tag{3.1}$$

où $\mathbf{S} = \mathbf{m} + \mathbf{U}_1\mathbf{y}_1$ est la composante dépendant seulement de l'identité du locuteur et $\mathbf{C}_r = \mathbf{U}_2\mathbf{y}_{2r} + \varepsilon_r$ est la composante dépendant seulement du canal, souvent nommée la composante du bruit. Ainsi, l'équation (3.1) peut se réécrire sous la forme suivante :

$$\mathbf{x}_r = \mathbf{m} + \mathbf{U}_1\mathbf{y}_1 + \mathbf{U}_2\mathbf{y}_{2r} + \varepsilon_r \tag{3.2}$$

[10] Nous considérons le domaine de la vérification du locuteur tel qu'il est proposé et évalué par l'institut américain *National Institute of Standards and Technology* (NIST, http://www.itl.nist.gov/iad/mig/tests/sre/).

où **m** est le vecteur moyen de dimension D estimé à partir d'une grande population de locuteurs, les matrices rectangulaires $\mathbf{U}_1$ de dimension $D \times d_1$ et $\mathbf{U}_2$ de dimension $D \times d_2$ sont des matrices de projection souvent nommées respectivement, la matrice des voix propres (*eigenvoices*) et la matrice des canaux propres (*eigenchannels*). Les vecteurs cachés $\mathbf{y}_1$ de dimension d_1 et $\mathbf{y}_2$ de dimension d_2 sont respectivement dénommés le vecteur des facteurs du locuteur (*speaker factors*) et le vecteur des facteurs du canal (*channel factors*). Enfin, ε_r est un vecteur de dimension D représentant le bruit résiduel, il suit une distribution gaussienne $N(\varepsilon_r, \mathbf{0}, \Lambda^{-1})$ de vecteur moyen **0** et de matrice de précision diagonale $\mathbf{\Lambda}$.

La distribution des vecteurs cachés $\mathbf{y}_1$ et $\mathbf{y}_2$ est souvent gaussienne standard (c.-à-d. $N(\mathbf{y}_\bullet, \mathbf{0}, \mathbf{I})$ de moyenne nulle **0** et de matrice de covariance identitaire **I**). Ce modèle génératif est dénommé PLDA gaussien. Le modèle PLDA à base de distribution à queue épaisse (*heavy-tailed*) est une autre variante de ce modèle génératif (Kenny, 2010a). Elle est caractérisée principalement par la distribution des vecteurs cachés qui suit une loi *t-student* plutôt qu'une loi gaussienne. Il est important de signaler qu'une estimation ponctuelle (*point estimate*) du vecteur $\mathbf{y}_1$ peut servir comme projection du i-vecteur $\mathbf{x}_r$ dans l'espace de dimension réduite du PLDA.

À vrai dire, la contribution réelle du terme $\mathbf{U}_2\mathbf{y}_{2r}$ de la composante du bruit $\mathbf{C}_r$ telle qu'elle est proposée dans le modèle (3.2) est négligeable dans le cas de la parole téléphonique. Ainsi, cette composante peut être réécrite tout simplement sous la forme : $\mathbf{C}_r = \varepsilon_r$, où le vecteur du bruit ε_r suit, dans ce cas-là, une gaussienne de vecteur moyen **0** et de matrice de précision pleine $\mathbf{\Lambda}$. Par la suite, nous avons aussi découvert que le terme de la composante du bruit cité ci-dessus n'est pas importante même dans le cas de la parole microphonique (Senoussaoui, *et al.*, 2011a) et ce, à condition de normaliser les i-vecteurs bruts via une Analyse discriminante linéaire (LDA). De plus, la projection des i-vecteurs normalisés par la LDA dans la sphère unitaire est nécessaire, obtenue par la normalisation à 1 de la norme euclidienne de ces i-vecteurs. Elle épargne la nécessité d'utilisation d'un modèle génératif complexe tel que le PLDA à base de la distribution *t-student* (Senoussaoui, *et al.*, 2011 b).

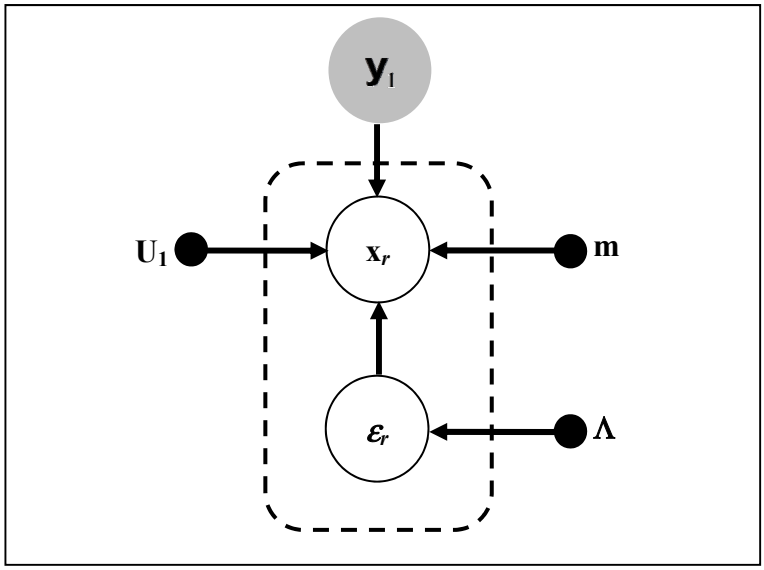

Figure 3.1 Graphe probabiliste du modèle PLDA donné par l'équation (3.3). Les points noirs représentent les paramètres du modèle, le cercle plein (arrière-plan gris) représente le vecteur caché, les deux cercles vides représentent respectivement un i-vecteur et un vecteur de bruit résiduel et enfin le cadre discontinu représente la répétition de la partie encadrée *R* fois.

Dans le cadre de cette thèse, nous nous focalisons sur le modèle PLDA gaussien simplifié, donné par l'équation suivante :

$$\mathbf{x}_r = \mathbf{m} + \mathbf{U}_1\mathbf{y}_1 + \varepsilon_r \tag{3.3}$$

Par ailleurs, nous pouvons ainsi simplifier les calculs subséquents par la réécriture de ce modèle comme suit :

$$\mathbf{x}_r = \mathbf{U}_1^+\mathbf{y}_1^+ + \varepsilon_r \tag{3.4}$$

où, $\mathbf{U}_1^+$ est la matrice augmentée de dimension $D \times (d_1+1)$ obtenue par la concaténation à droite du vecteur moyen $\mathbf{m}$ avec la matrice $\mathbf{U}_1$ et $\mathbf{y}_1^+$ est le vecteur augmenté de dimension (d_1+1) obtenu par la concaténation du chiffre 1 avec le vecteur caché $\mathbf{y}_1$.

Enfin, pour le lecteur intéressé, tous les détails nécessaires pour l'implémentation des autres versions du modèle PLDA sont dans l'article (Kenny, 2010a).

3.1.2 Apprentissage du modèle

L'apprentissage du modèle PLDA gaussien donné par l'équation (3.3) consiste à estimer le vecteur moyen $\mathbf{m}$, la matrice des voix propres $\mathbf{U}_1$ et la matrice de covariance $\mathbf{\Lambda}^{-1}$ de la composante du bruit (Prince, *et al.*, 2007)(Kenny, 2010a).

3.1.2.1 Distribution a posteriori des vecteurs cachés

Avant d'aborder en détail le processus d'apprentissage du modèle PLDA, il est essentiel de présenter le calcul de la distribution *a posteriori* $P(\mathbf{y}_1 \mid \mathbf{X})$du vecteur caché $\mathbf{y}_1$ étant donné l'ensemble $\mathbf{X} = \{\mathbf{x}_r\}_{r=1..R}$ de R enregistrements :

$$\begin{aligned} P(\mathbf{y}_1 \mid \mathbf{X}) &\equiv \ln P(\mathbf{X} \mid \mathbf{y}_1) + \ln P(\mathbf{y}_1) \\ &= \sum_{r=1}^{R} \ln N(\mathbf{x}_r, \mathbf{m} + \mathbf{U}_1\mathbf{y}_1, \Lambda^{-1}) + \ln N(\mathbf{y}_1, \mathbf{0}, \mathbf{I}) \\ &\equiv -\frac{1}{2}\sum_{r=1}^{R}(\mathbf{x}_r - \mathbf{m} + \mathbf{U}_1\mathbf{y}_1)' \Lambda (\mathbf{x}_r - \mathbf{m} + \mathbf{U}_1\mathbf{y}_1) - \frac{1}{2}\mathbf{y}_1'\mathbf{y}_1 \end{aligned} \tag{3.5}$$

Il est à noter que cette distribution *a posteriori* est calculée pour un locuteur donné ayant un ensemble de R enregistrements représentés par des i-vecteurs. L'équation de la distribution *a posteriori* (3.5) de $\mathbf{y}_1$ est quadratique par rapport à $\mathbf{y}_1$, ce qui implique que cette distribution est également une gaussienne tout comme sa distribution *a priori*. Le vecteur moyen et la matrice de covariance de cette distribution *a posteriori* peuvent être calculés en accumulant des statistiques d'ordre 1 et 2 comme suit :

$$\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_1) = (R\mathbf{U}_1'\Lambda\mathbf{U}_1 + \mathbf{I})^{-1} \tag{3.6}$$

$$\begin{aligned} \langle \mathbf{y}_1 \rangle &= (R\mathbf{U}_1'\Lambda\mathbf{U}_1 + \mathbf{I})^{-1}\mathbf{U}_1'\Lambda\sum_{r}^{R}(\mathbf{x}_r - \mathbf{m}) \\ &= \mathbf{Cov}(\mathbf{y}_1, \mathbf{y}_1)\mathbf{U}_1'\Lambda\sum_{r}^{R}(\mathbf{x}_r - \mathbf{m}) \end{aligned} \tag{3.7}$$

$$\langle \mathbf{y}_1\mathbf{y}_1 \rangle = \mathbf{Cov}(\mathbf{y}_1, \mathbf{y}_1) + \langle \mathbf{y}_1 \rangle \langle \mathbf{y}_1 \rangle' \tag{3.8}$$

3.1.2.2 Évaluation de la vraisemblance des données

Afin d'exploiter le modèle PLDA pour accomplir une tâche de classification, il est indispensable de fournir les formules nécessaires pour calculer la vraisemblance $P(\mathbf{X})$ d'un ensemble de données $\mathbf{X} = \{\mathbf{x}_r\}_{r=1..R}$ appartenant au même locuteur *l*. La quantité $P(\mathbf{X})$ est souvent connue dans le jargon bayésien sous l'appellation d'*Évidence* et donnée par l'intégrale suivante :

$$P(\mathbf{X}) = \int P(\mathbf{X}, \mathbf{y}_1)\, d\mathbf{y}_1 \tag{3.9}$$

Dans un cadre général (par exemple, une distribution « *heavy-tailed* » des vecteurs cachés), le calcul analytique direct de cette quantité est infaisable. Cependant, dans le cas du modèle gaussien, il existe une équation analytique directe pour calculer cette quantité (Prince, *et al.*, 2007). Les méthodes variationnelles sont, quant à elles, capables de fournir une bonne approximation via une limite inférieure ($\mathcal{L}$) de la quantité en question et dans le cas gaussien et dans le cas général (Kenny, 2010a) :

$$\mathcal{L} = E\left[\ln \frac{P(\mathbf{X}, \mathbf{y}_1)}{Q(\mathbf{y}_1)}\right] \tag{3.10}$$

où, *E* est l'espérance mathématique calculée en fonction du vecteur caché $\mathbf{y}_1$. La quantité $\mathcal{L}$ est toujours inférieure au logarithme de la vraisemblance réelle $\ln P(\mathbf{X})$, avec une égalité seulement et seulement si le dénominateur $Q(\mathbf{y}_1)$ est égal à la probabilité *a posteriori* réelle $P(\mathbf{y}_1 \mid \mathbf{X})$. La quantité $\mathcal{L}$ peut être réécrite comme suit :

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \tag{3.11}$$

où

$$\begin{aligned} \mathcal{L}_1 &= E\left[\ln P(\mathbf{X} \mid \mathbf{y}_1)\right] \\ \mathcal{L}_2 &= -KL\left(Q(\mathbf{y}_1) \middle\| P(\mathbf{y}_1)\right) \end{aligned} \tag{3.12}$$

KL est la divergence de Kullback-Leibler entre la distribution *a posteriori* approximée $Q(\mathbf{y}_1)$ du vecteur caché et la distribution *a priori* (gaussienne standard) $P(\mathbf{y}_1)$. L'évaluation de $\mathcal{L}_1$ est donnée comme suit :

$$\begin{aligned}\mathcal{L}_1 &= \sum_{r=1}^{R} E\left[\ln N\left(\mathbf{x}_r, \mathbf{U}_1^+\mathbf{y}_1^+, \Lambda^{-1}\right)\right] \\ &= \sum_{r=1}^{R}\left(\ln \frac{1}{(2\pi)^{D/2}\left|\Lambda^{-1}\right|^{1/2}}\right. \\ &\left. -\frac{1}{2}\left\langle\left(\mathbf{x}_r - \mathbf{U}_1^+\mathbf{y}_1^+\right)' \Lambda\left(\mathbf{x}_r - \mathbf{U}_1^+\mathbf{y}_1^+\right)\right\rangle\right) \\ &= \sum_{r=1}^{R}\left(\ln \frac{1}{(2\pi)^{D/2}\left|\Lambda^{-1}\right|^{1/2}} - \frac{1}{2}\left\langle\varepsilon_r'\Lambda\varepsilon\right\rangle\right)\end{aligned} \tag{3.13}$$

la quantité $\left\langle\varepsilon_r'\Lambda\varepsilon\right\rangle$ est donnée par :

$$\begin{aligned}\left\langle\varepsilon_r'\Lambda\varepsilon_r\right\rangle &= \mathrm{tr}\left(\left\langle\Lambda\varepsilon_r\varepsilon_r'\right\rangle\right) \\ &= \mathrm{tr}\left(\Lambda\ \mathrm{Cov}\left(\varepsilon_r, \varepsilon_r\right)\right) + \left\langle\varepsilon_r'\right\rangle\Lambda\left\langle\varepsilon_r\right\rangle \\ &= \mathrm{tr}\left(\mathbf{U}_1'\Lambda\mathbf{U}_1\ \mathrm{Cov}\left(\mathbf{y}_1, \mathbf{y}_1\right)\right) \\ &\quad +\left(\mathbf{x}_r - \mathbf{U}_1^+\left\langle\mathbf{y}_1\right\rangle\right)' \Lambda\left(\mathbf{x}_r - \mathbf{U}_1^+\left\langle\mathbf{y}_1\right\rangle\right).\end{aligned} \tag{3.14}$$

Enfin, la deuxième quantité $\mathcal{L}_2$ de notre limite inférieure est évaluée par l'équation suivante :

$$\begin{aligned}KL\left(Q(\mathbf{y}_1)\middle\|P(\mathbf{y}_1)\right) &= -\frac{d_1}{2} - \frac{1}{2}\ln\left|\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_1)\right| \\ &\quad +\frac{1}{2}\mathrm{tr}\left(\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_1) + \left\langle\mathbf{y}_1\right\rangle\left\langle\mathbf{y}_1'\right\rangle\right)\end{aligned} \tag{3.15}$$

En outre, les calculs de la limite inférieure présentés ci-dessus peuvent être accélérés par la diagonalisation simultanée des matrices Λ et $\mathbf{U}_1'\Lambda\mathbf{U}_1$. Les détails de l'estimation de la matrice de rotation réalisant cette diagonalisation sont présentés dans (Kenny, 2010a).

3.1.2.3 Algorithmes de mise à jour des paramètres du modèle

À ce stade, nous sommes en mesure de fournir l'algorithme d'apprentissage du modèle PLDA. D'une façon similaire au célèbre algorithme EM (Kenny, 2010a), cet algorithme se base sur la maximisation itérative de la vraisemblance (ML) des données. Tout comme l'apprentissage d'extracteur des i-vecteurs, un algorithme d'apprentissage à base de critère de la divergence minimale (MD) est également souhaitable dans le cas d'apprentissage du modèle PLDA. Dans ce qui suit, nous fournirons les équations mathématiques nécessaires pour faire les mises à jour itératives des paramètres $\{\mathbf{m}, \mathbf{U}_1, \Lambda^{-1}\}$. Ce processus itératif est amorcé d'une façon aléatoire avec des paramètres initiaux $\{\mathbf{m}^{(0)}, \mathbf{U}_1^{(0)}, \Lambda^{-1(0)}\}$. Nous assumons ainsi la présence d'un ensemble $\mathbf{X}$ de L locuteurs, chacun ayant $R(l)$ enregistrements d'apprentissage (avec un total de R enregistrements) représentés par des i-vecteurs, $\mathbf{X} = \{\mathbf{X}(l)\}_{l=1..L}$ et $\mathbf{X}(l) = \{\mathbf{x}_r(l)\}_{r=1..R(l)}$.

- ***Estimation ML***

L'objectif de cette méthode est de fournir des formules de mise à jour capables d'aboutir à des paramètres qui maximisent la vraisemblance totale des données d'apprentissage. La maximisation de la vraisemblance des données consiste à maximiser la quantité $\sum_l \mathcal{L}_1(l)$ de la limite inférieure, ce qui est nécessairement équivalent à minimiser l'équation suivante :

$$\sum_l \sum_{r=1}^{R(l)} \left\langle \left(\mathbf{x}_r(l) - \mathbf{U}_1^+ \mathbf{y}_1^+(l)\right)' \Lambda \left(\mathbf{x}_r(l) - \mathbf{U}_1^+ \mathbf{y}_1^+(l)\right) \right\rangle. \tag{3.16}$$

En mettant la dérivée par rapport à $\mathbf{U}_1^+$ de l'équation (3.16) égale à zéro, nous pouvons en déduire la formule de mise à jour de $\mathbf{U}_1^+$ comme suit :

$$\mathbf{U}_1^+ = \left[\sum_l \sum_{r=1}^{R(l)} \left\langle \mathbf{y}_1^+(l) \mathbf{y}_1^{+'}(l) \right\rangle\right]^{-1} \sum_l \sum_{r=1}^{R(l)} \mathbf{x}_r(l) \left\langle \mathbf{y}_1^+(l) \right\rangle. \tag{3.17}$$

Cette équation peut être résolue en calculant simplement la statistique suivante :

$$\left\langle \mathbf{y}_1^+ \mathbf{y}_1^{+\prime} \right\rangle = \begin{pmatrix} \mathrm{Cov}\left(\mathbf{y}_1^+, \mathbf{y}_1^+\right) + \left\langle \mathbf{y}_1^+ \right\rangle \left\langle \mathbf{y}_1^{+'} \right\rangle & \left\langle \mathbf{y}_1^{+'} \right\rangle \\ \left\langle \mathbf{y}_1^+ \right\rangle & 1 \end{pmatrix} \tag{3.18}$$

La formule de mise à jour de la matrice de bruit Λ^{-1} selon le critère du maximum de vraisemblance est :

$$\Lambda^{-1} = \frac{1}{R} \sum_l \sum_{r=1}^{R(l)} \left\langle \left(\mathbf{x}_r(l) - \mathbf{U}_1^+ \mathbf{y}_1^+(l)\right)\left(\mathbf{x}_r(l) - \mathbf{U}_1^+ \mathbf{y}_1^+(l)\right)' \right\rangle. \tag{3.19}$$

– ***Estimation MD***

Contrairement à l'objectif de l'estimation ML qui consiste à maximiser la partie de la vraisemblance approximée $\sum_l \mathcal{L}_1$ de la limite inférieure $\mathcal{L}$, l'objectif de l'estimation MD est de minimiser la deuxième partie de $\mathcal{L}$, à savoir, $\sum_l \mathcal{L}_2$ (Kenny, 2010a). Pour ce faire, une transformation doit être appliquée aux paramètres du modèle $\{\mathbf{m}, \mathbf{U}_1\}$ ainsi qu'au vecteur caché $\mathbf{y}_1$ de la manière suivante :

$$\bar{\mathbf{y}}_1(l) = \mathbf{B}\left(\mathbf{y}_1(l) - \mathbf{b}\right) \tag{3.20}$$

$$\hat{\mathbf{U}}_1 = \mathbf{U}_1 \mathbf{B}^{-1} \tag{3.21}$$

$$\hat{\mathbf{m}} = \mathbf{m} + \mathbf{U}_1 \mathbf{b}. \tag{3.22}$$

où **B** est une matrice de dimension $d_1 \times d_1$ et **b** est un vecteur de dimension d_1. Ces paramètres seront estimés de sorte qu'ils minimisent la somme des divergences suivante :

$$\sum_l \left(Q\left(\bar{\mathbf{y}}_1(l)\right) \middle\| P\left(\bar{\mathbf{y}}_1(l)\right) \right) \tag{3.23}$$

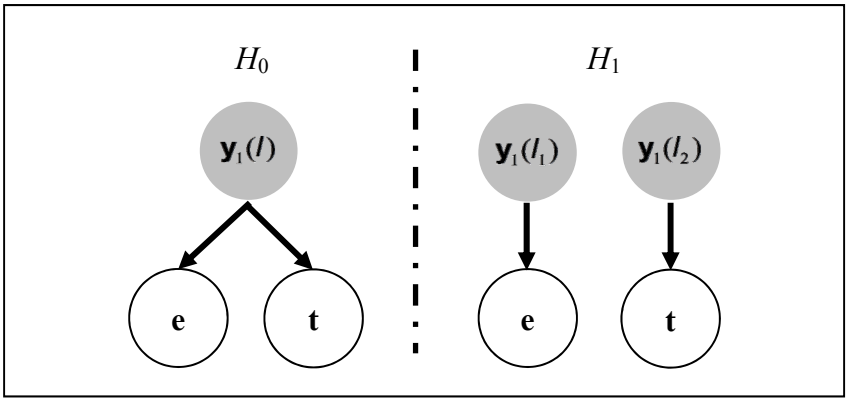

Figure 3.2 Graphe décrivant, dans le cadre du modèle PLDA, les deux hypothèses (H_0 , H_1) utilisées pour calculer le score de vérification en présence d'un i-vecteur d'enrôlement e et d'un autre i-vecteur pour le test t.

Les paramètres qui minimisent cette somme sont donnés par les équations suivantes :

$$\mathbf{b} = \frac{1}{L}\sum_{l}\langle \mathbf{y}_1(l)\rangle \tag{3.24}$$

$$\mathbf{B}^{-1} = \mathbf{K} \tag{3.25}$$

où **K** est une matrice triangulaire inférieure obtenue par la décomposition de Cholesky de la matrice suivante :

$$\frac{1}{L}\sum_{l}\langle \mathbf{y}_1(l)\mathbf{y}_1'(l)\rangle - \mathbf{b}\mathbf{b}'. \tag{3.26}$$

L'estimation par maximisation de la divergence (MD) a principalement deux caractéristiques majeures. D'abord, elle force la distribution du vecteur caché pour qu'elle soit une gaussienne standard, ensuite elle accélère la convergence de l'estimation des paramètres du modèle PLDA (Kenny, 2010a).

3.1.3 Vérification via le modèle PLDA

En présence d'un essai de vérification $x = \{\mathbf{e}, \mathbf{t}\}$ composé de deux enregistrements vocaux représentés par des i-vecteurs respectivement pour **e**nrôlement et pour **t**est, la tâche de la

vérification du locuteur via le modèle PLDA consiste à évaluer un score mesurant le degré de la ressemblance entre ces deux i-vecteurs. Comme nous l'avons introduit brièvement dans la Section 1.2.6, le rapport de vraisemblance LLR (voir l'équation 1.22) est le score calculé à travers un test d'hypothèses (voir Figure 3.2). À vrai dire, le LLR est le logarithme du rapport entre la probabilité que les deux i-vecteurs appartiennent au même locuteur (hypothèse H_0) et la probabilité qu'ils appartiennent à deux différents locuteurs (hypothèse H_1).

Les scores produits par le modèle PLDA de la manière précédemment décrite ne nécessitent aucune normalisation subséquente et dans le cas de la parole téléphonique (Kenny, 2010a) et dans le cas de la parole microphonique (Senoussaoui, *et al.*, 2011a).

3.2 Similarité angulaire du cosinus

Contrairement au modèle génératif décrit ci-dessus, la vérification du locuteur à base de la similarité angulaire du cosinus consiste simplement à calculer cette similarité entre les deux i-vecteurs composant un essai de vérification (voir Figure 3.3). Étant donné un essai de vérification $\boldsymbol{x} = \{\mathbf{e}, \mathbf{t}\}$, l'équation de la similarité angulaire du cosinus S s'écrit comme suit :

$$S(\mathbf{e}, \mathbf{t}) = \frac{\mathbf{e}' \cdot \mathbf{t}}{\|\mathbf{e}\| \|\mathbf{t}\|} \tag{3.27}$$

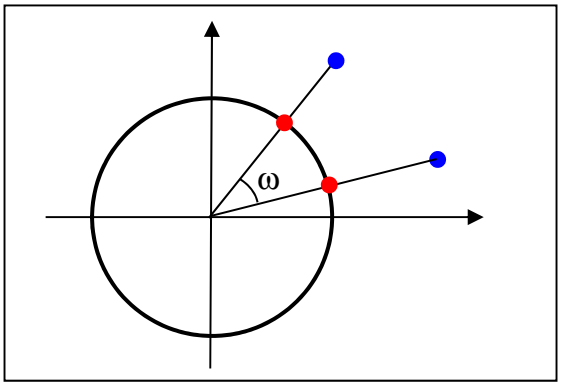

Figure 3.3 Exemple du calcul de la distance angulaire entre deux points. Les points rouges sont les projections sur la sphère des points bleus.

où ‖ ‖est la norme euclidienne d'un vecteur.

Afin que cette méthode fonctionne correctement, il est essentiel de normaliser les i-vecteurs de manière à minimiser la variabilité intralocuteur et de maximiser la variabilité interlocuteur.

3.2.1 Normalisation des i-vecteurs

La combinaison de l'analyse discriminante linéaire (Linear Discriminant Analysis, LDA) (Fisher, 1936) et la méthode de la normalisation avec la matrice de covariance intraclasse (Within Class Covariance Normalization, WCCN) (Hatch, *et al.*, 2006) est devenue un standard incontournable pour implémenter un système de la vérification du locuteur à base de la similarité angulaire.

3.2.1.1 Analyse discriminante linéaire (LDA)

L'Analyse discriminante linéaire ou bien l'analyse de Fisher (Fisher, 1936) est une méthode supervisée largement connue dans le domaine de la reconnaissance des formes. Elle est principalement destinée aux problèmes de la réduction de dimensionnalité. L'avantage principal de cette méthode est le fait de tenir compte des étiquettes des classes afin de former un sous-espace (c.-à-d. de dimension réduite) optimal en terme de la classification. À vrai dire, cet sous-espace minimise la variabilité intraclasse et maximise la variabilité interclasses. Pour ce faire, la LDA est basée sur l'optimisation de la fonction objective de Fisher $\mathbf{J}(\mathbf{u})$:

$$\mathbf{J}(\mathbf{u}) = \frac{\mathbf{u}'\mathbf{D}_b\mathbf{u}}{\mathbf{u}'\mathbf{D}_w\mathbf{u}} \tag{3.28}$$

où $\mathbf{D}_b$ est la matrice de dispersion interclasses, $\mathbf{D}_w$ est la matrice de dispersion intraclasse et $\mathbf{u}$ est un vecteur unitaire quelconque. La solution optimale de cette fonction est donnée par la résolution du problème généralisé des valeurs propres suivant :

$$\mathbf{D}_b\mathbf{u} = \lambda\mathbf{D}_w\mathbf{u} \tag{3.29}$$

cette fois-ci, **u** représente un vecteur propre et λ est sa valeur propre associée.

Les matrices de dispersions, interclasse et intraclasse, sont respectivement données par :

$$\mathbf{D}_b = \sum_{l}^{L} R(l)\left(\bar{\mathbf{x}}(l) - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}(l) - \bar{\mathbf{x}}\right)' \tag{3.30}$$

$$\mathbf{D}_w = \sum_{l}^{L} \sum_{r=1}^{R(l)} \left(\mathbf{x}_r(l) - \bar{\mathbf{x}}(l)\right)\left(\mathbf{x}_r(l) - \bar{\mathbf{x}}(l)\right)' \tag{3.31}$$

où L est le nombre total des classes (locuteurs dans notre cas) dans la base de données d'apprentissage, $R(l)$ est le nombre d'enregistrements $\{\mathbf{x}_r(l)\}_{r=1..R(l)}$ (i-vecteurs dans notre cas) du locuteur l, $\bar{\mathbf{x}}(l) = \frac{1}{R(l)} \sum_{r=1}^{R(l)} \mathbf{x}_r(l)$ est le vecteur moyen de toutes les données du locuteur l et $\bar{\mathbf{x}} = \frac{1}{R} \sum_{l=1}^{L} \sum_{r=1}^{R(l)} \mathbf{x}_r(l)$ est le vecteur moyen de toutes les données d'apprentissage.

Enfin, seuls les vecteurs propres ayant une valeur propre significative seront retenus pour construire la matrice de projection $\mathbf{P}_f$ de la LDA.

3.2.1.2 Normalisation via la matrice de covariance intraclasse (WCCN)

La normalisation de la variance des données en utilisant l'inverse de la matrice de covariance intraclasse est devenue une pratique courante dans le domaine de la reconnaissance du locuteur (Dehak, *et al.*, 2010)(Dehak, *et al.*, 2011 b)(Senoussaoui, *et al.*, 2013a). Cette méthode a été proposée pour la première fois par Hatch (Hatch, *et al.*, 2006) qui l'avait utilisée pour améliorer les performances d'un classificateur à vaste marge (Support Vector Machine, SVM). L'idée de base de cette normalisation est de pénaliser les axes contenant une forte variance intraclasse. Cette pénalisation est réalisée par une rotation de données via la matrice triangulaire inférieure $\mathbf{K}_w$ obtenue par la décomposition de Cholesky de l'inverse de la matrice de covariance intraclasse :

$$\hat{\mathbf{x}} = \mathbf{K}_w' \mathbf{x} \tag{3.32}$$

$\mathbf{C}_w = \frac{\mathbf{D}_w}{R}$ est la matrice de covariance intraclasse, R représente le nombre total des données d'apprentissage et $\mathbf{C}_w^{-1} = \mathbf{K}_w \mathbf{K}_w'$.

Afin d'être plus précis, la procédure standard de la vérification du locuteur via la similarité angulaire consiste à estimer la matrice $\mathbf{C}_w$ à partir des données préalablement projetées dans l'espace de dimension réduite de la LDA.

3.2.2 Vérification via la similarité du cosinus

Pour conclure, la procédure de normalisation des i-vecteurs dans le cadre d'un classificateur à base de similarité angulaire du cosinus peut être résumée par une reformulation de la similarité angulaire donnée par l'équation (3.27) de la manière suivante :

$$\widehat{S}(\mathbf{e},\mathbf{t}) = \frac{\left(\mathbf{P}_f'\mathbf{e}\right)' \mathbf{C}_w^{-1}\left(\mathbf{P}_f'\mathbf{t}\right)}{\sqrt{\left(\mathbf{P}_f'\mathbf{e}\right)' \mathbf{C}_w^{-1}\left(\mathbf{P}_f'\mathbf{e}\right)} \cdot \sqrt{\left(\mathbf{P}_f'\mathbf{t}\right)' \mathbf{C}_w^{-1}\left(\mathbf{P}_f'\mathbf{t}\right)}} \tag{3.33}$$

où $\mathbf{P}_f$ est la matrice de projection de la LDA et $\mathbf{C}_w$ est la matrice de covariance intraclasse estimée dans l'espace de dimension réduite de la LDA.

Contrairement au modèle PLDA, un système de vérification du locuteur à base de la similarité angulaire requiert une phase de normalisation des scores (voir Section 1.2.5). Nous fournirons plus de détails sur ce sujet dans le chapitre 5 de cette thèse.

CHAPITRE 4

INDÉPENDANCE DU CANAL

La disparité entre les conditions d'apprentissage et celles de l'exploitation d'un système de reconnaissance du locuteur reste, sans doute, la principale cause dégradante des performances de ces systèmes. Le changement radical du type du canal est effectivement un facteur important causant cette disparité (Sturim, *et al.*, 2007)(Jin, *et al.*, 2008). Dans ce contexte, nous entendons par le changement radical du canal, le passage du signal téléphonique au signal microphonique, ou vice versa. Nous pouvons également considérer le changement du type de microphone comme radical.

De nos jours, le besoin d'un traitement automatique des enregistrements audio ne cesse de s'accroitre. Ces fichiers audio sont souvent enregistrés au moyen de divers types de microphones et dans divers environnements acoustiques, ce qui constitue un véritable défi aux systèmes de traitement de la voix en général. Ainsi, il est devenu inévitable de poser cette question et d'en apporter les remèdes afin d'améliorer la robustesse de ces systèmes. Il est évident que ce problème peut être traité au niveau des vecteurs acoustiques en faisant appel à des méthodes de traitement du signal. Il est également possible de le traiter à un niveau supérieur tel que le niveau des i-vecteurs ou encore au niveau de la classification. Dans le cadre de cette thèse, nous nous focalisons sur le niveau des i-vecteurs et nous proposons deux solutions à ce problème.

4.1 Difficultés à surmonter

Outre que les facteurs habituels affectant la qualité du signal de la parole de manière générale, d'autres facteurs additionnels tels que la réverbération, l'écho et la position du locuteur par rapport au microphone peuvent compliquer davantage la reconnaissance dans le cas microphonique.

La représentation de la voix par des i-vecteurs est assez puissante, ainsi elle est devenue très populaire dans le domaine de la reconnaissance du locuteur (Dehak, *et al.* 2011a)(Kenny,

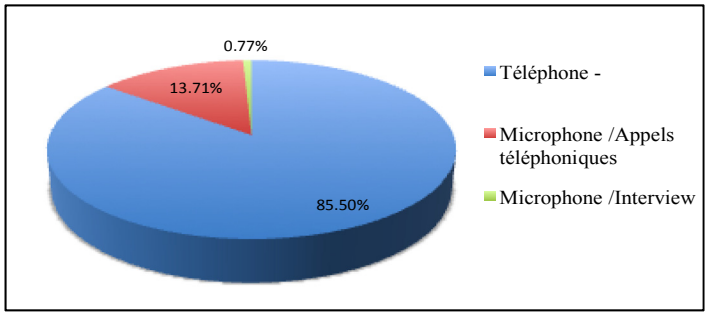

Figure 4.1 Représentation graphique des pourcentages des données de NIST réparties en fonction du type du canal (des enregistrements téléphoniques, des appels téléphoniques enregistrés par microphone et des enregistrements des interviews). Notons que le nombre total des enregistrements est de 41 706.

2010a). En effet, cette représentation a été largement investiguée avec succès notamment sur des données téléphoniques. L'une des clefs du succès de cette représentation est l'existence d'une énorme quantité de données téléphoniques fournies par NIST dans le cadre de ses évaluations périodiques de la reconnaissance du locuteur. Dans cette thèse, nous avons choisi de traiter le problème du changement du canal au niveau des i-vecteurs. Cependant, en faisant ce choix nous devons faire face à une pénurie de données microphoniques qui ne constituent que 14 % de l'ensemble total des données de NIST (voir Figure 4.1). Avec ce peu de données, il n'est même pas envisageable d'entrainer un extracteur d'i-vecteurs capable de capturer l'ensemble des variabilités de la parole notamment dans le cas microphonique. Afin de remédier à ce problème, nous proposons de profiter de l'énorme quantité de données téléphoniques pour entrainer un extracteur d'i-vecteurs convenant à la parole téléphonique et microphonique (Senoussaoui *et al.*, 2010).

4.2 Concaténation des matrices de la variabilité totale

Notre première contribution dans cette direction est de proposer un extracteur d'i-vecteurs indépendant du canal, dont la matrice **T** de dimension $CF \times D$ de la variabilité totale est obtenue par la concaténation de deux matrices. Ces deux matrices sont respectivement, une matrice de la variabilité totale $\mathbf{T}_{tel}$ de dimension $CF \times D_{tel}$, qui est estimée à partir des données téléphoniques, et la matrice $\mathbf{T}_{mic}$ de dimension $CF \times D_{mic}$, qui représente la matrice

de la variabilité totale estimée à partir des données microphoniques. Il est à noter que la dimension D de l'espace indépendant du canal est égale à la somme des dimensions D_{tel} et D_{mic} des deux sous-espaces dépendant du canal (téléphonique et microphonique).

Cette proposition était motivée par une architecture du modèle d'analyse conjointe des facteurs (JFA) implémentée avec succès spécialement pour traiter le problème des canaux croisés prescrit par NIST durant son évaluation de la reconnaissance du locuteur de l'année 2006 (Kenny *et al.*, 2008).

4.2.1 Définition du modèle

Dans le cas des i-vecteurs, un extracteur indépendant du canal (téléphone/microphone) est exprimé via un supervecteur **S** dépendant du locuteur et du canal de la manière suivante :

$$\mathbf{S} = \mathbf{m} + \mathbf{T}_{tel}\mathbf{x}_{tel} + \mathbf{T}_{mic}\mathbf{x}_{mic} \tag{4.1}$$

où $\mathbf{x}_{tel}$ et $\mathbf{x}_{mic}$ sont des vecteurs cachés dépendant respectivement du canal téléphonique et microphonique, **m** est le supervecteur moyen. Il est souvent recopié directement du supervecteur du modèle du monde.

4.2.1.1 Estimation des paramètres du modèle

La procédure d'entrainement de ce modèle (voir Figure 4.2) consiste à estimer les deux matrices de la variabilité totale $\mathbf{T}_{tel}$ et $\mathbf{T}_{mic}$. Cette procédure peut être résumée de la manière suivante.

Nous commençons par l'estimation de la matrice de la variabilité totale téléphonique $\mathbf{T}_{tel}$ selon le critère du maximum de vraisemblance similairement à celle décrite dans la Section 2.1. Ensuite, nous calculons le supervecteur $\mathbf{S}_r$ associé à chaque enregistrement de la base d'apprentissage microphonique en utilisant les paramètres téléphoniques comme suit :

$$\mathbf{S}_r = \mathbf{m} + \mathbf{T}_{tel}\mathbf{x}_{tel} \tag{4.2}$$

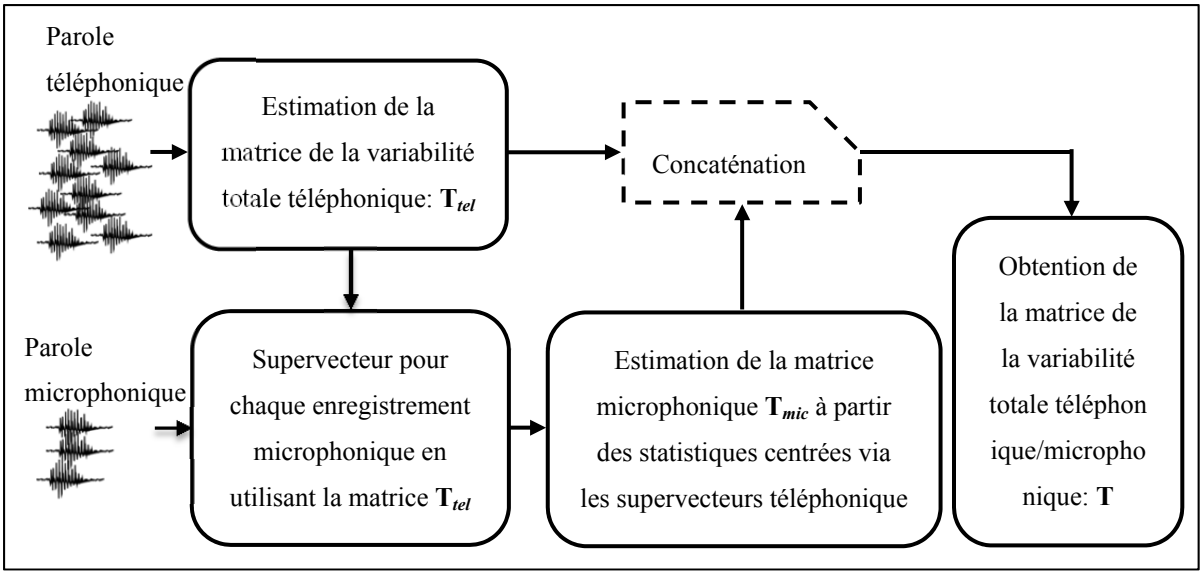

Figure 4.2 La procédure d'estimation par concaténation de la matrice de la variabilité totale indépendante du canal (téléphonique/microphonique).

Comme nous l'avons déjà défini, un supervecteur $\mathbf{S}_r$ est la concaténation de C vecteurs moyens $\mathbf{s}_r^{(c)}$ des composantes gaussiennes d'un modèle GMM. Nous utilisons ces vecteurs moyens du supervecteur d'un enregistrement pour centrer les statistiques de Baum-Welch de premier ordre $\mathbf{F}_r^{(c)}$ comme suit :

$$\mathbf{F}_r^{(c)} \leftarrow \left(\mathbf{F}_r^{(c)} - N_r^{(c)}\mathbf{s}_r^{(c)}\right) \tag{4.3}$$

où $N_r^{(c)}$ est le nombre des vecteurs acoustiques du $r^{ième}$ enregistrement alignés avec la $c^{ième}$ composante gaussienne.

Ces statistiques seront utilisées par la suite pour estimer la matrice de la variabilité totale microphonique $\mathbf{T}_{mic}$ selon la même procédure utilisée pour estimer $\mathbf{T}_{tel}$.

4.2.1.2 Extraction des i-vecteurs indépendants du canal

Après avoir estimé les deux matrices $\mathbf{T}_{tel}$ et $\mathbf{T}_{mic}$, l'extracteur des i-vecteurs indépendant du canal téléphonique/microphonique est ainsi obtenu par la concaténation de ces deux matrices comme suit :

$$\mathbf{S} = \mathbf{m} + \mathbf{T}\mathbf{x} \tag{4.4}$$

Comme nous l'avons déjà mentionné $\mathbf{T} = [\mathbf{T}_{tel}, \mathbf{T}_{mic}]$ est la matrice de la variabilité totale indépendante du canal obtenue par la concaténation des deux matrices $\mathbf{T}_{tel}$ et $\mathbf{T}_{mic}$ (voir Figure 4.2). Finalement, ce modèle est utilisé pour extraire les i-vecteurs selon la même procédure décrite dans la Section 2.1.5.

4.2.2 Expériences et résultats

Dans cette section, nous allons présenter quelques résultats expérimentaux de la vérification du locuteur obtenus par l'utilisation des i-vecteurs de dimension 600 qui sont extraits d'une manière indépendante du canal. Nous adoptons le système de la vérification du locuteur à base de la similarité angulaire du cosinus (voir Section 3.2). Les performances de ce système seront évaluées sur une liste d'essais fournie par NIST durant la campagne d'évaluation de l'année 2008. Cette liste dénommée « short2-short3 : *det*3 » ne contient que des segments de la parole microphonique et pour l'enrôlement et pour le test. Principalement, cette liste représente le défi le plus difficile du fait que tous les microphones d'enrôlement sont différents de ceux du test. Dans cette série d'expériences, seuls les résultats des locutrices seront présentés. Ceci est dû au fait que les listes d'essais de NIST sont séparées en fonction du genre du locuteur et que nos systèmes de vérification se comportent de la même manière sur les listes des hommes comme celles des femmes.

4.2.2.1 Détails d'implémentation

Avant d'exposer les expériences et les résultats, il est indispensable de présenter tous les détails nécessaires afin de reproduire conformément ces travaux :

- ***Modèle du monde (UBM)*** : nous avons entrainé deux modèles du monde dépendants du genre, dont chacun contenant $C = 2048$ gaussiennes. Ces modèles sont estimés à partir des données téléphoniques représentées par des vecteurs acoustiques de type MFCC de dimension $F = 60$ (19 coefficients MFCC + l'énergie + les premières et les secondes dérivées). Ces données sont les suivantes : LDC Switchboard II, Phases 2 et 3 ; Switchboard Cellulaire, Partie 1 et 2 et finalement les données « MIX » des évaluations de la reconnaissance du locuteur (SRE) de NIST des années 2004 et 2005.
- ***Extracteur des i-vecteurs*** : nous avons entrainé un extracteur d'i-vecteurs indépendant du type du canal (téléphonique/microphonique) de la même manière expliquée ci-dessus. D'abord, la matrice de la variabilité totale téléphonique, $\mathbf{T}_{tel}$ de dimension $CF \times 400$, est estimée à partir des mêmes données utilisées pour l'estimation de l'UBM. Nous estimons ensuite la matrice de la variabilité totale microphonique $\mathbf{T}_{mic}$, de dimension $CF \times 200$, à partir de toutes les données microphoniques de NIST (c.-à-d. NIST05, NIST06 et les données Interview de développement). Finalement, la matrice de la variabilité totale $\mathbf{T}$ (de dimension $CF \times 600$) indépendante des canaux (téléphonique/microphonique) est obtenue par la concaténation de la matrice $\mathbf{T}_{tel}$ et $\mathbf{T}_{mic}$.
- ***Méthodes de classification*** : dans cette première série d'expériences, nous allons adopter un classificateur à base de la similarité angulaire du cosinus (SAC) pour implémenter notre système de vérification. Nous allons également présenter les résultats obtenus via des machines à vaste marge (SVM) à base d'un noyau du cosinus (*Cosine Kernel*) (Dehak *et al,*. 2011a)(Senoussaoui, *et al.*, 2010). Les détails de l'aspect théorique des SVM peuvent être consultés dans l'article (Vapnik, 1998). De plus, les résultats du modèle d'Analyse conjointe des facteurs (JFA) seront aussi présentés à des fins de comparaison. Enfin, nous fournirons des résultats supplémentaires obtenus dans (Kenny,

2010a) en utilisant le même extracteur d'i-vecteurs proposé ci-dessus et un modèle génératif de type PLDA pour la classification.

- ***Compensation des effets du canal*** : Afin de compenser les variabilités indésirables de la parole propagées jusqu'aux i-vecteurs, nous avons fait appel à la combinaison de l'analyse discriminante linéaire (LDA) et à la normalisation via l'inverse de la matrice de covariance intraclasse (WCCN) décrite en Section 3.2.1. Selon le type des données utilisées et la façon d'estimation des matrices de LDA et de WCCN, nous proposons trois stratégies d'estimations de ces matrices :
  1. *Stratégie 1* : l'estimation des matrices de la normalisation LDA et WCCN se fait à partir des données microphoniques.
  2. *Stratégie 2* : l'estimation des matrices de la normalisation LDA et WCCN se fait à partir d'un mélange de données microphoniques et téléphoniques.
  3. *Stratégie 3* : dans cette dernière stratégie, nous estimons séparément les paramètres à partir des données téléphoniques et microphoniques. Ensuite, nous combinons les matrices téléphoniques avec leurs homologues microphoniques. La combinaison est exprimée par une somme pondérée, comme dans le cas de la matrice interclasses $\mathbf{D}_b$:

$$\mathbf{D}_b = \alpha^{(tel)}\mathbf{D}_b^{(tel)} + \alpha^{(mic)}\mathbf{D}_b^{(mic)} \tag{4.5}$$

où, $\mathbf{D}_b^{(tel)}$ est la matrice interclasses téléphonique, $\mathbf{D}_b^{(mic)}$ est la matrice interclasses microphonique, $\alpha^{(tel)}$ et $\alpha^{(mic)}$ sont respectivement les facteurs de pondérations téléphoniques et microphoniques.

- ***Normalisation des scores*** : dans cette étude nous avons adopté la combinaison de la méthode de normalisation *z-norm* avec celle de *t-norm* (voir Section 1.2.5). Cette combinaison est dénommée *zt-norm*. Dans la pratique nous avons utilisé 200 modèles d'imposteurs pour la *t-norm* et 1007 observations des imposteurs pour la *z-norm*.
- ***Analyse conjointe des facteurs (JFA)*** : nous ne nous intéressons pas au modèle JFA (Kenny, *et al.*, 2008) dans cette étude. Toutefois, nous présenterons les résultats de ce modèle à des fins de comparaison avec nos résultats. La configuration adoptée du modèle JFA est la suivante : 300 facteurs du locuteur « *speaker factors* » et 100 facteurs du canal

« *channel factors* » qui sont estimés à partir des mêmes données téléphoniques utilisées dans l'entrainement du modèle du monde UBM. 100 facteurs du canal additionnels sont estimés à partir de l'ensemble de toutes les données microphoniques de NIST (microphone et interview).

4.2.2.2 Résultats et discussions

Tableau 4.1 Les résultats obtenus par la similarité angulaire du cosinus (SAC) et les machines à vaste marge (SVM) testées sur la tâche « *short2-short3 : det3* » de NIST SRE 2008 (locuteurs femmes).

		EER(%)	MinDCF	dim
Stratégie 1	SAC	6.1	0.034	600
	SVM	**5.1**	**0.024**	600
Stratégie 2	SAC	7.1	0.040	400
	SVM	**6.4**	**0.033**	400
Stratégie 3	SAC	5.4	0.022	400
	SVM	**4.5**	**0.022**	400
JFA		**3.9**	**0.021**	-

En observant les résultats présentés dans le Tableau 4.1, nous tirons quelques conclusions comme suit : d'abord, les résultats obtenus via les SVM sont meilleurs que ceux de la SAC. Les résultats obtenus par les i-vecteurs sont comparables à ceux obtenus par le modèle JFA (voir lignes 7 et 8 du Tableau 4.1). Finalement, la troisième stratégie (c.-à-d. la somme pondérée des matrices téléphoniques et microphoniques) d'estimation des matrices de LDA et de WCCN semble la plus efficace (voir lignes 6 et 7 du Tableau 4.1).

Nous exposons également dans cette section (voir Tableau 4.2), des résultats obtenus par Kenny (Kenny, 2010a) en utilisant le même extracteur des i-vecteurs défini par l'équation (4.1). Il est à noter que les résultats de la parole microphonique (c.-à-d. *det*1, *det*4 et *det*5) sont obtenus via un modèle PLDA complet (c.-à-d. le modèle donné par l'équation 3.2) ; ce modèle nécessite une normalisation des scores de type *s-norm* (voir Section 1.2.5.3). Cependant, les résultats de la parole téléphonique (*det*7) sont obtenus via le modèle PLDA

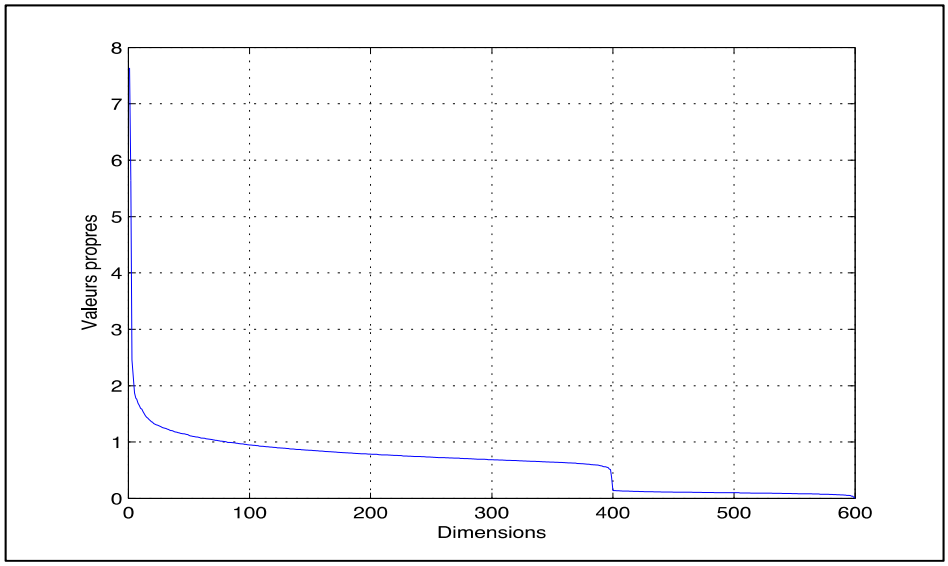

Figure 4.3 Graphe des valeurs propres de la matrice de covariance des i-vecteurs (obtenus par la concaténation des matrices de la variabilité totale) des données d'entrainement de NIST.

simplifié (voir équation 3.3) et sans avoir besoin d'aucune méthode de normalisation. Les deux modèles cités ci-dessus sont à base de la distribution *t-student* des vecteurs cachés. Pour plus de résultats et de détails théoriques et techniques, veuillez vous référer à (Kenny, 2010a).

Tableau 4.2 Résultats de Kenny (Kenny, 2010a) obtenus pour les tâches short2-short3 : *det*1, *det*4, *det*5 (parole microphonique) et *det*7 (parole téléphonique) de NIST SRE 2008 (locuteurs femmes).

	EER(%)	MinDCF
*det*1	3.4	0.017
*det*4	3.1	0.018
*det*5	3.8	0.020
*det*7	2.2	0.010

4.2.2.3 PLDA pour la réduction de dimensionnalité

Après avoir obtenu ces résultats encourageants, nous avons décidé d'investiguer davantage l'idée de la concaténation des matrices de la variabilité totale. Durant l'évaluation de la

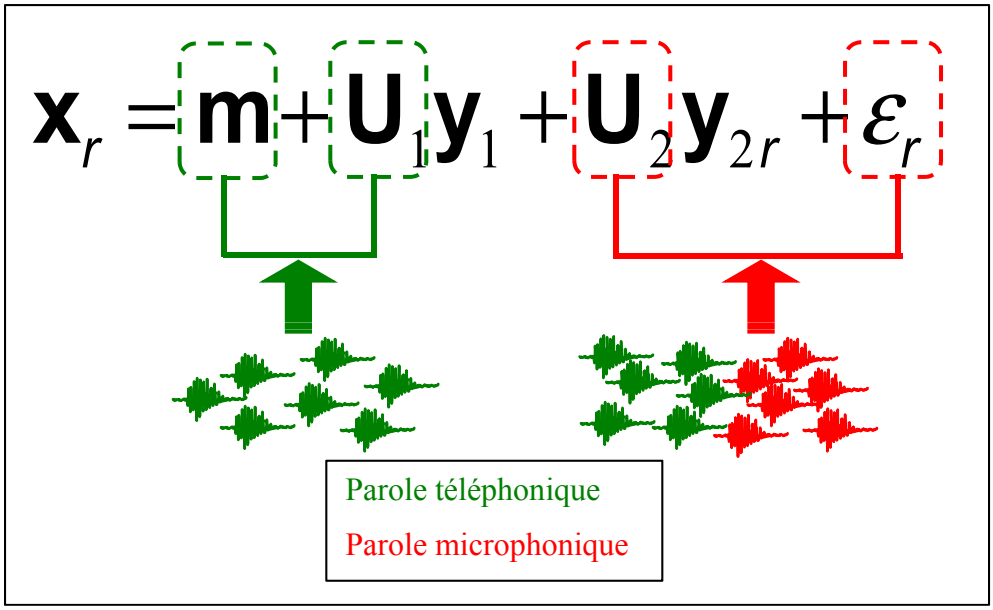

Figure 4.4 Procédure d'entrainement des paramètres du modèle PLDA destiné à la réduction de la dimensionnalité.

reconnaissance du locuteur de NIST 2010, le groupe du MIT (*Massachusetts Institute of Technology*) a proposé d'utiliser le modèle génératif PLDA pour réduire la dimensionnalité de nos i-vecteurs extraits d'une manière indépendante du canal comme décrit ci-dessus (Dehak, *et al.*, 2011a)(Sturim, *et al.*, 2011). Dans le cas de cet extracteur d'i-vecteurs (indépendant du canal), cette réduction de dimensionnalité est destinée à réduire à la fois la variabilité indésirable et l'hétérogénéité des composantes des i-vecteurs qui est engendrées principalement par la concaténation des deux sous-espaces (téléphonique/microphonique) de natures différentes. Cette hétérogénéité est clairement perçue par l'irrégularité de la courbe des valeurs propres de la matrice de covariance des données d'apprentissage de NIST (voir la dimension 400 à la Figure 4.3).

Nous avons testé cette nouvelle architecture avec des données microphoniques fournies par NIST durant la campagne d'évaluation de 2010 (liste élargie des tests de la tâche « coreext-coreext : *det* 2 »). Les résultats sont présentés au Tableau 4.3. Dans cette série d'expériences, nous avons utilisé les mêmes i-vecteurs de dimension 600 décrits ci-dessus. Tout d'abord, la dimensionnalité des i-vecteurs a été réduite de 600 à 400 en utilisant le modèle PLDA à base de distribution *t-student* donné par l'équation (3.2). La meilleure stratégie d'estimation des

paramètres de ce modèle PLDA destiné à la réduction de dimensionnalité est la suivante (voir Figure 4.4) : les paramètres **m** et $\mathbf{U}_1$ représentant la composante dépendante du locuteur sont estimés à partir des données téléphoniques. Cependant, le reste des paramètres de ce modèle (c.-à-d. $\mathbf{U}_2$ et ε_r) représentant la composante dépendante du canal sont estimés à partir d'un mélange de données téléphoniques et microphoniques. Ensuite, ces i-vecteurs de dimension réduite sont utilisés pour entrainer un nouveau modèle PLDA simplifié (voir équation 3.3) à base de la distribution *t-student* qui servira cette fois-ci à classifier la voix des locuteurs lors de la vérification.

4.2.2.4 Résultats et discussions

Tableau 4.3 Résultats de la tâche *coreext-coreext* : *det*2 de NIST SRE 2010 (locuteurs femmes) obtenus via un modèle PLDA simplifié à base de distribution t-student. La dimensionnalité des i-vecteurs indépendants du canal est réduite via PLDA.

	EER(%)	MinDCF_08[11]	MinDCF_10[11]	dim
Sans *s-norm*	3.94	0.217	0.686	400
Avec *s-norm*	**3.25**	**0.172**	**0.554**	400
Résultats de LPT	3.61	-	0.577	-

Les résultats de Loquendo Politecnico di Torino (LPT) sont également fournis dans la dernière ligne du Tableau 4.3 à des fins de comparaison. En observant le Tableau 4.3, nous pouvons résumer les avantages de la réduction supervisée de dimensionnalité comme suit. Les résultats obtenus sont légèrement meilleurs que ceux de l'état de l'art obtenus par la fusion des scores des décisions de **huit différents systèmes** (résultats de LPT). De plus, ces résultats sont obtenus via le modèle PLDA simplifié et sans aucune normalisation de scores à l'inverse de ceux obtenus dans la série des expériences précédentes.

[11] MinDCF_08 représente le minimum de la DCF calculé à partir des paramètres du coût fixées par NIST durant SRE 2008 et MinDCF_10 est celui de SRE 2010.

4.3 Entrainement à partir des données regroupées

Après avoir découvert la nécessité de la réduction supervisée de la dimensionnalité des i-vecteurs dans le processus de la vérification, nous avons décidé d'explorer l'idée de l'entrainement d'un extracteur d'i-vecteurs à partir d'un seul ensemble regroupant les données téléphoniques et microphoniques. Encouragés par la réussite de l'idée de concaténation des matrices des canaux propres (Kenny, *et al.*, 2008) et découragés par le caractère non équilibré de l'ensemble des données téléphoniques et microphoniques (voir Figure 4.1), nous avions totalement écarté cette idée au départ.

4.3.1 LDA pour la réduction de dimensionnalité

L'objectif principal de la réduction supervisée de la dimensionnalité des i-vecteurs est d'éliminer les axes d'espace des données maximisant la variabilité intraclasse tout en gardant uniquement ceux qui maximisent la variabilité interclasses. Cette réduction supervisée de la dimensionnalité réalisée à travers le modèle génératif PLDA dans les expériences précédentes peut également s'accomplir via la méthode d'analyse discriminante linéaire (LDA) ordinaire. Par analogie avec la stratégie d'entrainement des paramètres du modèle génératif PLDA destiné à la réduction de dimensionnalité, l'estimation des matrices de dispersions interclasse et intraclasse (voir Section 3.2.1.1) de LDA se fait de la manière suivante :

- La matrice de dispersion interclasse modélise uniquement la variabilité interlocuteur. Ainsi, une estimation de cette matrice à partir d'un ensemble de données homogène et suffisamment large est souhaitable. Dans le cadre de NIST, l'ensemble des données téléphoniques (voir Figure 4.1) est parfaitement adéquat pour l'estimation de cette matrice.
- La matrice de dispersion intraclasse, quant à elle, modélise l'ensemble des variabilités indésirables à la tâche de la vérification du locuteur. L'ensemble d'apprentissage le plus convenable pour estimer cette matrice est celui qui contient des données provenant de

différents types de canaux de transmissions. Donc, l'ensemble regroupant les données téléphoniques et microphoniques de NIST est idéal pour estimer cette matrice.

4.3.2 Expériences et résultats

Afin de valider l'efficacité de ce nouvel extracteur d'i-vecteurs, nous avons estimé la matrice de la variabilité totale, de dimension $CF \times 800$, de cet extracteur à partir de l'ensemble regroupant toutes les données d'apprentissage de NIST (voir Figure 4.1). L'entrainement de cet extracteur se fait de la manière habituelle présentée dans le Chapitre 2.

Une LDA est entrainée par la suite de la même manière expliquée dans la section précédente afin de réduire la dimensionnalité des i-vecteurs de 800 à 200. De plus, la normalisation à 1 de la norme euclidienne de ces i-vecteurs (voir Sections 2.2.1 et 3.1.1) est souhaitable pour rendre leur distribution proche d'une loi normale multidimensionnelle (Garcia-Romero, 2011).

Tableau 4.4 Résultats de la tâche *coreext-coreext* : *det*2 et *det*5 de NIST SRE 2010 obtenus via un modèle PLDA simplifié à base de distribution gaussienne (locutrices).

	EER(%)	MinDCF_08[12]	MinDCF_10[12]
*det*2	3.8	0.190	0.543
*det*5	2.4	0.124	0.387

Les i-vecteurs de dimension 200 et de norme euclidienne égale à 1 sont ensuite utilisés pour entrainer un modèle PLDA simplifié à base de distribution gaussienne (voir l'équation 3.3). Ce modèle est testé par la suite sur la tâche « *coreext-coreext : det*2 » impliquant la parole microphonique provenant de deux différents microphones. Nous avons également testé ce modèle sur la tâche « *coreext-coreext : det*5 » impliquant la parole téléphonique. Les résultats de ces expériences sont fournis par le Tableau 4.4.

[12] MinDCF_08 représente le minimum de la DCF calculé à partir des paramètres du coût fixées par NIST durant SRE 2008 et MinDCF_10 est celui de SRE 2010.

Les résultats de la parole microphonique (*det*2) sont comparables à ceux de l'expérience précédente (voir Tableau 4.3) obtenus via un modèle PLDA à base de la distribution *t-student*, et ce, notamment en ce qui concerne la métrique MinDCF_10. De bons résultats sont également obtenus dans le cas de la parole téléphonique (*det*5). En fait, ces résultats sont considérés comme les résultats de l'état de l'art actuel des systèmes de la vérification du locuteur évalués selon les protocoles de NIST.

CHAPITRE 5

INDÉPENDANCE DU GENRE

Dans le chapitre précédent, nous avons détaillé notre recherche qui porte sur le problème du changement radical du type de canal de transmission dans le cadre des systèmes de la vérification du locuteur à base des i-vecteurs. Afin d'apporter des solutions à cette question, nous avons choisi d'intervenir au niveau de l'extraction des i-vecteurs. Ce choix est principalement motivé par le manque de données microphoniques dans le contexte expérimental de NIST. Dans ce chapitre, nous explorerons une autre piste aussi importante que celle de l'indépendance du canal et qui nous mènera à améliorer davantage la robustesse des systèmes de la vérification du locuteur face à des conditions d'exploitation réelles. Cette piste consiste à traiter le problème de l'indépendance du genre du locuteur, toujours, dans le cadre des systèmes de vérification à base des i-vecteurs.

Les listes des essais de test, prescrites par NIST dans le cadre de ses évaluations, ont trois caractéristiques principales. En effet, pour un essai de test donné, i) ces listes fournissent l'information concernant le genre des locuteurs impliqués, ii) le type de canal des segments vocaux et iii) elles ne permettent pas de croisement de genres (c.-à-d. un essai dont le segment d'enrôlement et celui du test proviennent de locuteurs de genre différent).

Influencés par les campagnes d'évaluation de NIST et ses protocoles expérimentaux, les chercheurs du domaine de la reconnaissance du locuteur ont tendance à concevoir des systèmes de reconnaissance du locuteur essentiellement dépendants du genre du locuteur et du type du canal. Dans la vie réelle, ces distinctions sont quasi incontrôlables, ce qui pose un grand défi aux systèmes de la reconnaissance du locuteur exploités dans des milieux hors des laboratoires.

Contrairement aux solutions proposées au niveau d'extracteur des i-vecteurs pour remédier au problème de la dépendance du canal de transmission, nous proposons dans ce chapitre des solutions au niveau des classificateurs pour pallier le problème de la dépendance au genre.

Ceci est principalement dû à l'existence quasiment équiprobable des vastes quantités de données d'apprentissage de NIST pour les hommes et pour les femmes.

Nos solutions proposées sont destinées à deux types de système de vérification du locuteur considérés comme ceux de l'état de l'art actuel. Le premier est basé sur le modèle génératif PLDA (voir Section 1.2.3.4) (Senoussaoui, *et al.*, 2011a), le deuxième est à base de la similarité angulaire du cosinus (voir Section 1.2.3.4) (Senoussaoui, *et al.*, 2013a).

5.1 Modèle génératif indépendant du genre

5.1.1 PLDA indépendant du genre (PLDA-IG)

La façon la plus simple et naïve de concevoir un système de vérification du locuteur à base du modèle génératif PLDA est de l'estimer simplement à partir d'un seul ensemble regroupant les données des deux genres (homme et femme). La modélisation de ces données de nature multimodale (au moins bimodale) par cette simple stratégie est sous-optimale du point de vue théorique. Cependant, nous expérimenterons cette stratégie afin d'établir un système de référence.

5.1.2 Mélange des modèles PLDA (PLDA-M)

De par sa nature probabiliste, le modèle génératif PLDA permet une grande flexibilité au niveau de la modélisation et du calcul des scores. De plus, l'ensemble regroupant les données des femmes et celles des hommes peut être vu comme un ensemble bimodal de données. Par conséquent, une modélisation de cet ensemble par un mélange de deux modèles génératifs de type PLDA dépendant du genre (PLDA-DG) (c.-à-d., qui représentent respectivement les femmes et les hommes) s'avère la plus adéquate dans le cas des systèmes de la vérification du locuteur à base de PLDA.

5.1.2.1 Définition du modèle du mélange

La définition élémentaire de la tâche de vérification du locuteur, comme déjà présentée mainte fois, consiste à comparer deux segments vocaux (chacun est représenté par un i-vecteur de dimension *D* dans notre cas) d'un essai $x = \{\mathbf{e}, \mathbf{t}\}$ de test afin de juger si ces deux segments appartiennent au même locuteur ou non.

En réalité, un modèle PLDA ne tient pas compte des vraies complexités de la production, transmission et traitement de la parole. Cependant, il considère que les i-vecteurs représentant cette parole sont simplement produits par des processus aléatoires. Ainsi, ce modèle considère que chaque i-vecteur de l'essai $x = \{\mathbf{e}, \mathbf{t}\}$ est produit par les modèles PLDA (voir Équation 3.3) comme suit :

$$\begin{aligned} \mathbf{e} &= \mathbf{m} + \mathbf{U}\mathbf{y}^{(\mathbf{e})} + \varepsilon^{(\mathbf{e})} \\ \mathbf{t} &= \mathbf{m} + \mathbf{U}\mathbf{y}^{(\mathbf{t})} + \varepsilon^{(\mathbf{t})} \end{aligned} \tag{5.1}$$

où les vecteurs cachés $\mathbf{y}^{(\mathbf{e})}$ et $\mathbf{y}^{(\mathbf{t})}$ de dimension *d* représentent l'identité des locuteurs. Ils sont générés par la distribution interlocuteur multidimensionnelle continue et les vecteurs de bruits $\varepsilon^{(\mathbf{e})}$ et $\varepsilon^{(\mathbf{t})}$ de dimension *D* sont générés par la distribution intralocuteur multidimensionnelle continue. Dans le cas d'un essai cible (c.-à-d., les deux segments appartiennent au même locuteur), les vecteurs cachés sont égaux $\mathbf{y}^{(\mathbf{e})} = \mathbf{y}^{(\mathbf{t})}$ et ils diffèrent dans le cas inverse.

Outre que les variables cachées continues $\mathbf{y}^{(\mathbf{e})}$ et $\mathbf{y}^{(\mathbf{t})}$ du modèle PLDA, nous avons deux autres variables cachées discrètes $\mathcal{C}$ et $\bar{\mathcal{C}}$ signifiant respectivement un essai cible et un essai non-cible. Ces deux variables sont nos variables d'intérêt qui devront être inférées pour chaque essai présenté au modèle PLDA.

L'implémentation du mélange des modèles PLDA implique l'entrainement de deux modèles, le premier à partir des données des hommes et le deuxième à partir des données des femmes.

Les paramètres de ces deux modèles sont estimés de la manière habituelle et sans aucune modification (voir Section 3.1.2), ils seront utilisés lors de l'étape du test.

5.1.2.2 Modélisation du genre du locuteur

Afin de modéliser le genre du locuteur, nous devons ajouter deux autres variables cachées discrètes g_e et g_t qui prennent la valeur h pour *homme* et f pour *femme*. Évidemment, pour un essai cible, la valeur de ces deux variables est égale, ce qui n'est pas nécessairement le cas pour un essai non-cible. Des valeurs probabilistes *a priori* devront être associées à ces variables selon le type d'essai comme suit :

$$\begin{aligned} P_{hh} &= P(hh \mid C), \quad P_{ff} = P(ff \mid C) \\ Q_{hh} &= P(hh \mid \bar{C}), \quad Q_{ff} = P(ff \mid \bar{C}) \\ Q_{hf} &= P(hf \mid \bar{C}), \quad Q_{fh} = P(fh \mid \bar{C}) \end{aligned} \tag{5.2}$$

où par exemple hf désigne l'évènement : $g_e = h$ et $g_t = f$. Ces valeurs devront satisfaire les conditions suivantes : $P_{hh} + P_{ff} = 1$ et $Q_{hh} + Q_{ff} + Q_{hf} + Q_{fh} = 1$. Dans la pratique, l'utilisateur du système de vérification du locuteur a la possibilité de fixer ces valeurs selon ses besoins. Dans cette recherche nous fixons ces valeurs d'une façon équiprobable.

5.1.2.3 Calcul de score

Jusqu'alors, nous n'avons pas encore vu clairement la différence entre le mélange de PLDA et le modèle PLDA ordinaire. Principalement, cette différence est perçue au niveau du module qui calcule les scores de vérification. En effet, ce module est la seule composante du modèle PLDA-DG ordinaire (c.-à-d. dépendant du genre) qui doit subir une simple modification selon les lois de la probabilité, et ce, afin de le rendre indépendant du genre du locuteur.

Dans le but de réaliser le calcul des scores indépendamment du genre, nous avons besoin d'abord d'évaluer, d'une manière dépendante du genre, les vraisemblances suivantes :

$$P(\mathbf{e},\mathbf{t}\,|\,\hbar\hbar,\mathcal{C})$$
$$P(\mathbf{e},\mathbf{t}\,|\,ff,\mathcal{C}) \tag{5.3}$$

pour les essais cibles et

$$P(\mathbf{e},\mathbf{t}\,|\,\hbar\hbar,\bar{\mathcal{C}}) = P(\mathbf{e}\,|\,\hbar)P(\mathbf{t}\,|\,\hbar)$$
$$P(\mathbf{e},\mathbf{t}\,|\,ff,\bar{\mathcal{C}}) = P(\mathbf{e}\,|\,f)P(\mathbf{t}\,|\,f) \tag{5.4}$$

pour les essais non-cibles. Ces vraisemblances sont calculées de la manière habituelle (voir Section 3.1.2.2) en utilisant deux modèles PLDA dépendants du genre c.à.d. un modèle pour les locuteurs et un autre modèle pour les locutrices.

Nous pouvons également arranger ces vraisemblances d'une manière à obtenir des rapports de vraisemblances (LLR) comme suit :

$$R_{\hbar} = \frac{P(\mathbf{e},\mathbf{t}\,|\,\hbar\hbar,\mathcal{C})}{P(\mathbf{e}\,|\,\hbar)P(\mathbf{t}\,|\,\hbar)} \tag{5.5}$$

$$R_{f} = \frac{P(\mathbf{e},\mathbf{t}\,|\,ff,\mathcal{C})}{P(\mathbf{e}\,|\,f)P(\mathbf{t}\,|\,f)} \tag{5.6}$$

$$G_i = \frac{P(i\,|\,\hbar)}{P(i\,|\,f)}, \quad i \in \{\mathbf{e},\mathbf{t}\}. \tag{5.7}$$

Notons que les LLR données par les équations (5.5) et (5.6) représentent les LLR ordinaires utilisés pour la vérification dépendante du genre. Le logarithme du rapport G_i et $i \in \{\mathbf{e},\mathbf{t}\}$ peut être utilisé comme score d'un détecteur du genre, d'ailleurs, on a atteint une erreur de détection égale à 2 % lorsqu'on l'avait testé sur les données téléphoniques de NIST SRE2010.

Soit $\tilde{R}$ le score LLR indépendant du genre du locuteur. Ce score est obtenu par marginalisation (sommation) des variables du genre comme suit :

$$\tilde{R} = \frac{P(\mathbf{e},\mathbf{t}\,|\,\mathcal{C})}{P(\mathbf{e},\mathbf{t}\,|\,\bar{\mathcal{C}})} = \frac{P_{\hbar} p(\mathbf{e},\mathbf{t}\,|\,\hbar\hbar,\mathcal{C}) + P_{f} p(\mathbf{e},\mathbf{t}\,|\,ff,\mathcal{C})}{\sum_{g_e,g_t} Q_{g_e g_t} P(\mathbf{e}\,|\,g_e) P(\mathbf{t}\,|\,g_t)} \tag{5.8}$$

où le dénominateur est formé par la somme de quatre termes.

Finalement, le score $\tilde{R}$ indépendant du genre peut également se réécrire en fonction des LLR dépendants du genre et des détecteurs du genre (voir les équations 5.5, 5.6 et 5.7) comme suit :

$$\tilde{R} = \frac{P_{\hbar}}{Q_{\hbar\hbar}} S_{\hbar} R_{\hbar} + \frac{P_{f}}{Q_{ff}} S_{f} R_{f} \tag{5.9}$$

où

$$S_{\hbar} = \frac{Q_{\hbar\hbar} G_e G_t}{Q_{\hbar\hbar} G_e G_t + Q_{f\hbar} G_t + Q_{\hbar f} G_e + Q_{ff}} \tag{5.10}$$

$$S_{f} = \frac{Q_{ff}}{Q_{ff} + Q_{\hbar f} G_e + Q_{f\hbar} G_t + Q_{\hbar\hbar} G_e G_t} \tag{5.11}$$

5.1.2.4 Les essais à genre croisé

À première vue, ce problème semble facile, car la discrimination entre deux locuteurs de genres différents est plus facile que s'ils étaient du même genre. Toutefois, ce n'est pas toujours le cas. Prenant l'exemple d'un système traditionnel GMM-UBM de vérification dépendant du genre du locuteur qui a besoin des méthodes de la normalisation des scores telle que la *t-norm*. Supposons qu'on fait face à un essai de vérification non-cible composé de deux locutrices. Notre détecteur de genre à base d'une décision stricte (*hard decision*) a déterminé à tort que les locuteurs sont plutôt des hommes. En conséquence, nous sélectionnons les cohortes des imposteurs hommes pour la *t-normalisation*, produisant ainsi

un score de vérification très élevé. Donc, notre système va conclure par erreur que l'essai en question est plutôt un essai cible.

Notons qu'à nos jours il n'existe aucune liste officielle contenant des essais à genre croisé. Ainsi, la seule façon d'expérimenter ce problème reste de créer sa propre liste d'essais. Afin de réaliser nos expériences nous avons établi notre propre liste d'essais à genre croisé de la manière suivante. Nous avons gardé tous les essais cibles (hommes/femmes) de la liste téléphonique « *coreext-coreext* : *det*5 » de NIST et nous avons créé une liste d'essais non-cibles dont tous les essais sont à genre croisé. Le nombre des essais à genre croisé dans cette deuxième liste est égal au nombre des essais non-cibles de la liste originale de NIST « *coreext-coreext* : *det*5 ».

Outre que la liste d'essais à genre croisé, nous avons également fixé un protocole expérimental afin de nous permettre de tester adéquatement les performances de nos systèmes vis-à-vis de ce problème. Ce protocole consiste à utiliser le seuil de décision optimal (c.-à-d. le seuil correspondant au MinDCF) estimé à partir de la liste d'essais officielle de NIST « *coreext-coreext* : *det*5 » afin de calculer un DCF (*Detection Cost Function*) à partir des scores des tests prescrits sur la liste d'essais à genre croisé. Par conséquent, cette DCF est espérée être inférieure au minimum de DCF (MinDCF) calculé sur la liste officielle, car tous les essais non-cibles sont des essais à genre croisé.

5.1.3 Expérimentations

Dans cette section, nous présenterons des résultats expérimentaux de la vérification du locuteur via le modèle PLDA-DG dépendant du genre, le modèle indépendant du genre PLDA-IG ainsi que le modèle de mélange PLDA-M. Dans le but de réaliser ces expériences, un extracteur des i-vecteurs de dimension $D = 800$ a été entrainé d'une manière indépendante du genre et du canal à partir d'un ensemble regroupant les données téléphoniques et microphoniques des deux genres (voir Section 4.3). Les performances de ces systèmes seront testées sur les listes élargies (*extended lists*) des essais des cinq premières conditions de la tâche principale (*coreext-coreext*) fournies par NIST durant la campagne d'évaluation (SRE)

de 2010, à savoir, *det*1, *det*2, *det*3, *det*4 et *det*5. Les mesures des performances de NIST (EER, MinDCF_08[12] et MINDCF_10[12]) sont utilisées pour l'évaluation des performances de tous les systèmes (voir Section 1.2.7.1).

5.1.3.1 Détails d'implémentation

Avant de présenter les expériences et les résultats, il est indispensable de présenter tous les détails nécessaires pour reproduire ces travaux :

- ***Modèle du monde (UBM)*** : nous avons entrainé un seul modèle du monde indépendant du genre qui contient $C = 2048$ gaussiennes. Ce modèle est estimé à partir des données représentées par des vecteurs acoustiques de type MFCC de dimension $F = 60$ (19 coefficients MFCC + l'énergie + les premières et les secondes dérivées). Ces données sont les suivantes : LDC Switchboard II, Phases 2 et 3 ; Switchboard Cellulaire, Parties 1 et 2 et finalement les données « MIX » des évaluations de la reconnaissance du locuteur (SRE) de NIST des années 2004 et 2005.
- ***Extracteur des i-vecteurs*** : nous avons entrainé un extracteur d'i-vecteurs indépendant et du type du canal (téléphonique/microphonique) et du genre du locuteur (homme/femme) de la manière habituelle. La matrice de la variabilité totale **T** de dimension $CF \times 800$, est estimée à partir les données des deux genres suivantes : LDC releases of Switchboard II, Phases 2 et 3; Switchboard Cellulaire, Parties 1 et 2 ; les données de Fisher, NIST SRE2004 et SRE2005 (c.-à-d. la parole téléphonique) ainsi que les données de NIST 05, 06, 08 et les données microphoniques de développement (c.-à-d. toutes les données microphoniques de NIST).
- ***Projection via LDA*** : la projection des i-vecteurs dans l'espace de dimension réduite de la LDA (estimée de la même manière décrite dans la Section 4.3.1 et suivie par la normalisation à **1** de la norme euclidienne) est une étape indispensable pour le fonctionnement du modèle PLDA à base de la distribution gaussienne (Senoussaoui, *et al.*, 2011a)(Garcia-Romero, 2011). Les matrices de dispersions de LDA ont été entrainées à partir des mêmes ensembles de données, regroupant hommes et femmes, utilisés pour entrainer l'extracteur des i-vecteurs (c.-à-d. les données téléphoniques et

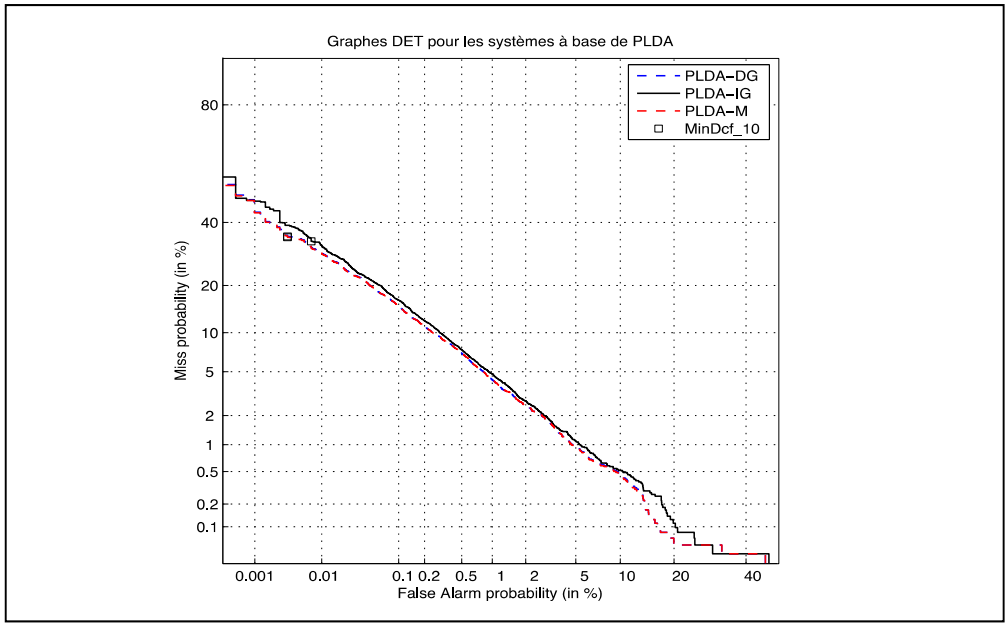

Figure 5.1 Courbes DET des systèmes à base de PLDA testés sur les listes téléphoniques de NIST (det5).

microphoniques de NIST) à l'exception des données de Fisher. Enfin, la dimension réduite optimale $d = 200$ a été déterminée empiriquement.

– ***Modèles PLDA*** : Trois modèles PLDA ont été entrainés pour nos expériences, à savoir, un modèle indépendant du genre (PLDA-IG) et deux modèles PLDA dépendants du genre. Tous les modèles ont été entrainés à partir des mêmes ensembles de données utilisés pour entrainer l'extracteur des i-vecteurs (c.-à-d. les données téléphoniques et microphoniques de NIST à l'exception des données de Fisher). Pour les trois modèles, les matrices des voix propres $\mathbf{U}_1$ correspondantes sont de dimension $d \times d$ (c.-à-d. 200×200). Le mélange des modèles PLDA est mis en œuvre en combinant les deux modèles dépendants du genre (c.-à-d. les modèles masculin et féminin) tel qu'expliqué dans la Section 5.1.2.3.

5.1.3.2 Résultats et discutions

Nous commençons cette série d'expériences par tester les trois configurations des systèmes à base des modèles PLDA sur la liste de la condition du test *det*5 (c.-à-d. téléphone/téléphone).

Tableau 5.1 Résultats des différents systèmes PLDA testés sur la liste det5 (téléphone/téléphone) de NIST SRE 2010.

		EER (%)	MinDCF_08	MinDCF_10
Femmes	PLDA-DG	**2.47**	**0.124**	**0.387**
	PLDA-IG	2.75	0.133	0.415
	PLDA-M	**2.46**	**0.124**	**0.388**
Hommes	PLDA-DG	**1.81**	**0.096**	**0.320**
	PLDA-IG	2.00	0.112	0.386
	PLDA-M	**1.81**	**0.096**	**0.322**

Les résultats présentés par le Tableau 5.1 révèlent clairement l'efficacité du modèle de mélange PLDA-M face au problème de l'indépendance du genre du locuteur. En effet, les résultats obtenus par le modèle de mélange PLDA-M sans aucune information sur le genre des locuteurs impliqués dans les essais sont essentiellement les mêmes que ceux obtenus par le modèle dépendant du genre PLDA-DG, où l'information sur le genre est toujours fournie. De plus, ce comportement est perçu tout au long de la courbe *DET* (voir Figure 5.1). Par ailleurs, les performances du système naïf indépendant du genre PLDA-IG étaient constamment inférieures par rapport au reste des systèmes. Ceci confirme notre hypothèse de départ stipulant que le modèle PLDA-IG est assez simple pour pouvoir modéliser la nature bimodale des données.

La deuxième expérience de cette série vise à étudier le comportement de notre système à base du mélange PLDA-M vis-à-vis des essais à genre croisé. Afin de réaliser ces expériences, nous avons suivi le protocole expliqué à la Section 5.1.2.4.

Tableau 5.2 Résultats du système PLDA-M testé sur la liste des essais à genre croisé (téléphone/téléphone).

		EER (%)	DCF_08	DCF_10
Femmes	Liste de NIST	2.46	0.124	0.388
	Liste d. g. croisé	**0.28**	**0.082**	**0.332**
Hommes	Liste de NIST	1.81	0.096	0.322
	Liste d. g. croisé	**0.45**	**0.064**	**0.300**

Tous les résultats obtenus sur la liste des essais à genre croisé sont meilleurs que ceux obtenus sur la liste officielle de NIST. C'est un accomplissement qui s'harmonise parfaitement avec ce qui est espéré dans le protocole expérimental proposé dans la Section 5.1.2.4.

Tableau 5.3 Résultats des différents systèmes PLDA testés sur la liste det2 (interview/interview) de NIST SRE 2010.

		EER (%)	MinDCF_08	MinDCF_10
Femmes	PLDA-DG	3.86	0.190	0.543
	PLDA-IG	**3.80**	**0.187**	**0.536**
	PLDA-M	3.87	0.190	0.541
Hommes	PLDA-DG	**2.02**	**0.097**	**0.363**
	PLDA-IG	2.11	0.098	0.397
	PLDA-M	**2.03**	**0.097**	**0.365**

Nous avons expérimenté avec succès l'efficacité du modèle de mélange PLDA-M par rapport au modèle indépendant du genre simple (PLDA-IG), et ce, lorsqu'ils étaient testés vis-à-vis de la parole téléphonique (*det*5). Dans ce qui suit, nous présenterons les résultats montrant les comportements de ces systèmes dans le cas de l'utilisation de la parole microphonique (*det*1, *det*2 et *det*3) et de la parole mixte (*det*4), tout en commençant par la condition *det*2 qui est en fait la plus difficile (voir Tableau 1.2).

Nous pouvons tirer deux conclusions principales des résultats de la condition *det*2. Tout d'abord, nous pouvons constater que les résultats du PLDA-M (voir lignes 4 et 7 du Tableau 5.3) sont semblables aux résultats du PLDA-DG (voir lignes 2 et 5 du Tableau 5.3). Ainsi, nous concluons que le modèle de mélange est efficace aussi face aux données microphoniques. D'autre part, la comparaison entre les résultats des locutrices du modèle PLDA-M avec ceux du modèle simple PLDA-IG (voir ligne 3 du Tableau 5.3) révèle une anomalie (voir Annexe III pour de plus amples informations concernant les intervalles de confiance des résultats). Étant donné que les résultats du modèle PLDA-IG sont même légèrement meilleurs que ceux du modèle dépendant du genre PLDA-DG, nous favorisons l'hypothèse stipulant que ce comportement est probablement dû à une erreur expérimentale.

Tableau 5.4 Résultats des différents systèmes PLDA testés sur les listes det1, det3 et det4, regroupant les locuteurs hommes et femmes, de NIST SRE 2010.

		EER (%)	MinDCF_08	MinDCF_10
*det*1	PLDA-DG	1.58	0.070	0.246
	PLDA-IG	**1.44**	0.071	0.262
	PLDA-M	1.58	0.070	0.246
*det*3	PLDA-DG	2.68	0.126	0.402
	PLDA-IG	**2.57**	**0.124**	0.439
	PLDA-M	2.68	0.125	0.397
*det*4	PLDA-DG	2.90	0.128	0.385
	PLDA-IG	3.05	0.133	0.403
	PLDA-M	2.90	0.129	0.384

Afin de faciliter la comparaison aux lecteurs, les résultats des locuteurs hommes et femmes seront regroupés pour les conditions *det*1, *det*3 et *det*4 de NIST SRE 2010 (voir Tableau 5.4). Ces résultats confirment de nouveau l'efficacité de la vérification indépendante du genre (c.-à-d. dans l'absence totale d'information sur le genre des locuteurs) via le modèle de mélange. Les cellules en caractères gras du Tableau 5.4 montrent que nous avons encore une fois la même anomalie observée que dans le cas de *det*2 (voir Tableau 5.3). Or, il s'avère clair, par le caractère intermittent d'apparition de ce comportement, que ces résultats représentent des cas isolés et non pas une tendance générale.

5.2 Similarité angulaire indépendante du genre

Contrairement au modèle PLDA, un classificateur à base de similarité angulaire du cosinus (SAC) n'est pas de nature probabiliste. Par conséquent, la conception d'un classificateur indépendant du genre du locuteur dans ce cas-là, n'est pas aussi simple. Nous proposons donc une combinaison des similarités angulaires pondérée par les probabilités d'un détecteur du genre afin d'implémenter un module de calcul des similarités angulaires indépendant du genre.

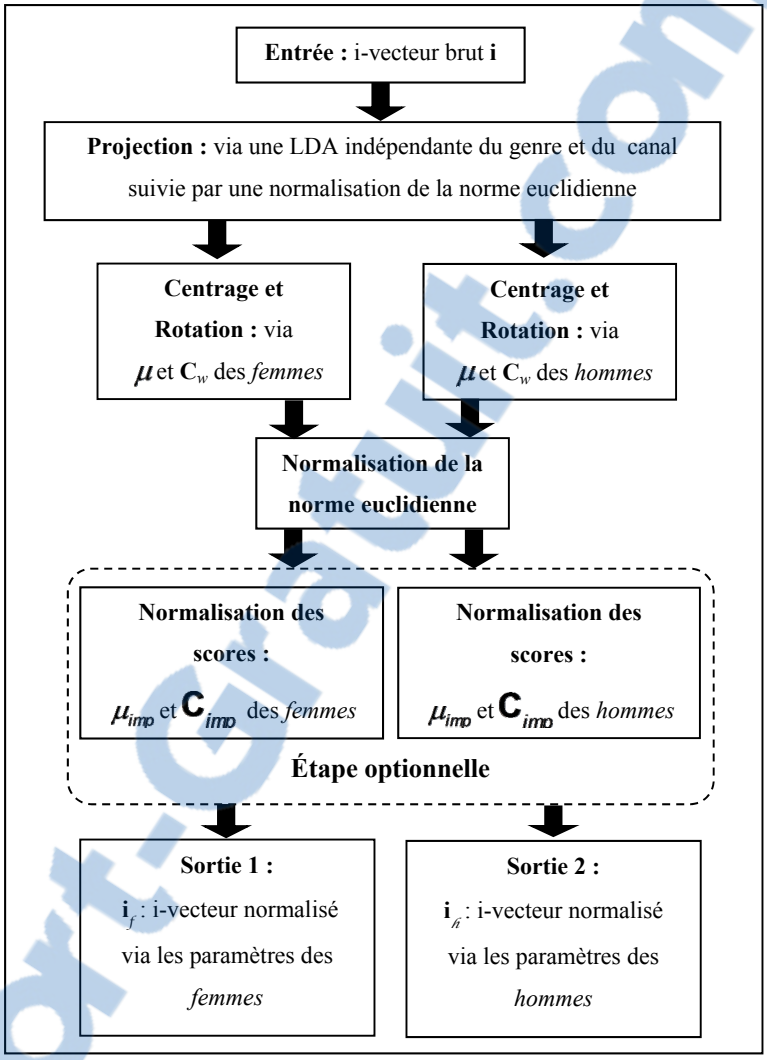

Figure 5.2 Représentation graphique du processus de la normalisation indépendante du genre d'un i-vecteur **i** dans le cadre d'un classificateur à base de la similarité angulaire du cosinus.

5.2.1 La SAC dépendante du genre (SAC-DG)

Avant d'aborder le problème de l'indépendance du genre, il est préférable de fournir la procédure détaillée de la vérification à base de la similarité angulaire du cosinus, présentée brièvement dans la Section 3.2. Un score de vérification à base de la similarité entre deux i-vecteurs d'un essai de test $x = \{\mathbf{e}, \mathbf{t}\}$ donnée par l'équation (3.27), peut se réécrire par un

simple produit scalaire de ces deux i-vecteurs après avoir normalisé à **1** leurs normes euclidiennes comme suit :

$$score(\mathbf{e},\mathbf{t}) = \mathbf{e}' \cdot \mathbf{t} \tag{5.12}$$

où **e** et **t** sont normalisés comme suit :

$$\mathbf{e} = \frac{\mathbf{e}}{\|\mathbf{e}\|} \quad \text{et} \quad \mathbf{t} = \frac{\mathbf{t}}{\|\mathbf{t}\|} \tag{5.13}$$

En réalité, la normalisation de la norme euclidienne est effectuée après une série de transformations visant à maximiser la variabilité interlocuteur et à minimiser la variabilité intralocuteur (voir Figure 5.2) (Dehak, *et al.*, 2010).

5.2.1.1 Compensation des effets du canal

Comme évoquée mainte fois antérieurement, la procédure de la compensation des effets indésirables du canal la plus répondue dans le cadre de la vérification à base de la SAC est celle combinant la LDA et la WCCN (Dehak, *et al.*, 2010). À vrai dire, cette procédure implique aussi une opération de centrage des i-vecteurs en utilisant un i-vecteur moyen μ estimé à partir d'une importante quantité de données d'apprentissage. L'opération du centrage des données est effectuée entre les normalisations via la LDA et la WCCN.

5.2.1.2 Normalisation des scores

Il est courant que les scores de vérification calculés par la similarité angulaire nécessitent une normalisation de type *zt-norm* ou bien de type *s-norm* (voir Session 1.2.5), et ce, afin de réduire la variabilité indésirable qui s'y est propagée dés le début du processus de la vérification. Traditionnellement, l'estimation des paramètres de la normalisation ainsi que l'application des équations normalisant les scores se font souvent en ligne (c.-à-d. au moment de la vérification). Une nouvelle reformulation des équations de la normalisation *zt-norm*

adaptée au cas de la similarité angulaire du cosinus a été proposée par Dehak (Dehak, *et al.*, 2010). La nouvelle équation de la normalisation *zt-norm* est donnée par :

$$zt\_score(\mathbf{e},\mathbf{t}) = \frac{(\mathbf{e}-\mu_{imp})'\cdot(\mathbf{t}-\mu_{imp})}{\left\|\mathbf{C}_{imp}\cdot\mathbf{e}\right\|\cdot\left\|\mathbf{C}_{imp}\cdot\mathbf{t}\right\|} = \left(\frac{(\mathbf{e}-\mu_{imp})}{\left\|\mathbf{C}_{imp}\cdot e\right\|}\right)'\cdot\left(\frac{(\mathbf{t}-\mu_{imp})}{\left\|\mathbf{C}_{imp}\cdot t\right\|}\right) \tag{5.14}$$

où μ_{imp} est un vecteurs moyen estimé d'une cohorte des i-vecteurs d'imposteurs, normalisés selon la procédure illustrée sur la Figure 5.2, et $\mathbf{C}_{imp}$ est une matrice triangulaire issue d'une décomposition de Cholesky de la matrice de covariance, estimée à partir de la même cohorte des i-vecteurs d'imposteurs.

Nous pouvons également utiliser ces deux mêmes paramètres (μ_{imp} et $\mathbf{C}_{imp}$) afin d'évaluer une nouvelle reformulation de la normalisation *s-norm* (voir l'équation originale 1.21) des scores comme suit :

$$s\_score(\mathbf{e},\mathbf{t}) = \left((\mathbf{e}-\mu_{imp})'\cdot\frac{\mathbf{t}}{\left\|\mathbf{C}_{imp}\cdot\mathbf{t}\right\|}\right) + \left((\mathbf{t}-\mu_{imp})'\cdot\frac{\mathbf{e}}{\left\|\mathbf{C}_{imp}\cdot\mathbf{e}\right\|}\right) \tag{5.15}$$

5.2.2 La SAC indépendante du genre

Notre stratégie d'implémentation d'un système de vérification du locuteur indépendant du genre repose, comme nous l'avons évoqué ci-dessus, sur une combinaison des scores dépendants du genre pondérée par les probabilités d'un détecteur de genre. Ainsi, il est fort intéressant de concevoir un détecteur de genre robuste. Dans la section suivante, nous présenterons les détails d'implémentation du détecteur de genre opérant dans l'espace des i-vecteurs que nous avons adopté dans ce travail.

5.2.2.1 Détecteur du genre d'un locuteur

Le rôle principal d'un détecteur du genre d'un locuteur est d'affecter un segment vocal en se basant sur les caractéristiques de la voix du locuteur dans l'une des deux classes possibles (c.-à-d. *homme* ou *femme*). Donc, il est tout simplement possible d'implémenter un détecteur de genre par l'entrainement d'un classificateur à deux classes. Dans le cadre de ce travail, nous proposons d'explorer l'idée de la modélisation de chaque classe représentant l'un des deux genres par une distribution gaussienne des i-vecteurs comme suit :

$$P(\mathbf{i} \mid f) \propto N(\mathbf{i}, \mu_f, \Sigma_f) \tag{5.16}$$

$$P(\mathbf{i} \mid h) \propto N(\mathbf{i}, \mu_h, \Sigma_h) \tag{5.17}$$

Habituellement, les paramètres (vecteur moyen μ_g et matrice de covariance $\mathbf{\Sigma}_g$) des distributions de chaque classe de genre sont estimés à partir des ensembles dépendants du genre. Or, l'utilisation directe du i-vecteur moyen μ et de la matrice de covariance intraclasse $\mathbf{C}_w$ dépendants du genre (voir la Figure 5.2) comme paramètres permet d'obtenir des résultats similaires à ceux obtenus dans le cas habituel. En procédant ainsi, nous épargnons une étape supplémentaire d'apprentissage de notre détecteur de genre.

Tableau 5.5 Les résultats (l'erreur de détection (*Err*) et le nombre d'observations (*N. Obs*)) des détecteurs de genre à base de distribution gaussienne testés sur les données de NIST SRE2010 (*det*1... *det*5). Les paramètres des gaussiennes sont estimés à partir de l'ensemble des données d'apprentissage de NIST (téléphoniques et microphoniques).

	Hommes		Femmes		***Moyen/Total***	
	Err (%)	*N. Obs*	*Err (%)*	*N. Obs*	*Err (%)*	*N. Obs*
Données de *det1*	1.08	1108	2.33	1283	1.70	2391
Données de *det2*	1.33	3882	2.31	4498	1.82	8380
Données de *det3*	1.19	1510	2.36	1777	1.77	3287
Données de *det4*	1.16	1548	2.18	1692	1.67	3240
Données de *det5*	1.91	2294	2.11	2740	2.01	5034
Moyen/Total	**1.33**	**10 342**	**2.25**	**11 990**	**1.79**	**22 332**

Une fois testé sur les données téléphoniques (données de la tâche *det*5) de NIST SRE 2010, ce détecteur de genre a permis l'obtention de ~2 % du taux d'erreur de détection (voir Tableau 5.5 pour plus de résultats).

5.2.2.2 La SAC indépendante du genre (SAC-IG)

Tout comme dans le cas de PLDA-GI, la façon la plus simple (naïve) de concevoir un système de vérification du locuteur, à base de la similarité angulaire du cosinus indépendante du genre (SAC-IG), consiste à estimer les paramètres (c.-à-d. les paramètres de normalisation des i-vecteur et des scores) de ce modèle à partir d'un seul ensemble regroupant les données des deux genres. D'après notre expérience avec le PLDA-GI et de point de vue théorique, cette stratégie est loin d'être optimale pour le problème de l'indépendance du genre. Cependant, nous l'adoptons dans le but d'avoir un système de référence utilisé pour des fins de comparaison.

5.2.2.3 Combinaison des SAC (SAC-C)

Supposons d'abord que nous avons produit pour chaque i-vecteur d'un essai de vérification $x = \{\mathbf{e}, \mathbf{t}\}$ deux i-vecteurs ($\{\mathbf{e}_h, \mathbf{e}_f\}$ et $\{\mathbf{t}_h, \mathbf{t}_f\}$) normalisés respectivement via des paramètres des hommes et ceux des femmes (voir Figure 5.2). Le *zt-score* indépendant du genre peut être obtenu par une simple somme des scores dépendants du genre. Cette somme est pondérée par les probabilités du détecteur de genre (voir les équations 5.16 et 5.17) comme suit :

$$\begin{aligned} zt\_score(\mathbf{e},\mathbf{t}) &= \left(p(\mathbf{e}\,|\,h)\cdot\mathbf{e}'_h \cdot p(\mathbf{t}\,|\,h)\cdot t_h\right) + \\ &\quad \left(p(\mathbf{e}\,|\,f)\cdot\mathbf{e}'_f \cdot p(\mathbf{t}\,|\,f)\cdot\mathbf{t}_f\right) \\ &= \left(\lambda_{hh}\cdot(\mathbf{e}'_h\cdot\mathbf{t}_h)\right) + \left(\lambda_{ff}\cdot(\mathbf{e}'_f\cdot\mathbf{t}_f)\right) \end{aligned} \tag{5.18}$$

où les facteurs de pondération sont calculés par les équations suivantes :

$$\lambda_{hh} = P(\mathbf{e}|h)P(\mathbf{e}|h) \tag{5.19}$$

$$\lambda_{ff} = P(\mathbf{e}|f)P(\mathbf{e}|f) \tag{5.20}$$

Le score *s-normalisé* peut être également obtenu de la même manière décrite ci-dessus.

5.2.3 Expérimentations

Comme dans le cas du modèle génératif, nous expérimentons les systèmes à base de similarité angulaire du cosinus (SAC) sur les mêmes listes d'essais (c.-à-d. les listes officielles de NIST et la liste des essais à genre croisé).

5.2.3.1 Détails d'implémentation

Dans cette série d'expériences, nous reprenons exactement les mêmes configurations du modèle du monde (UBM), de l'extracteur des i-vecteurs et de la LDA adoptées dans l'expérimentation du modèle génératif PLDA (voir Section 5.1.3.1).

- ***Matrice de covariance intraclasse et vecteur moyen*** : dans le cas des systèmes de vérification du locuteur à base de la SAC, le processus de la normalisation des i-vecteurs consiste à centrer les i-vecteurs déjà projetés dans l'espace de la LDA via un vecteur moyen μ et les faire pivoter par la WCCN (c.-à-d. une rotation via la décomposition de Cholesky de l'inverse de la matrice de covariance intraclasse $\mathbf{C}_w$) (voir Figure 5.2). Contrairement à l'UBM, à l'extracteur des i-vecteurs et à la LDA, ces deux paramètres sont estimés d'une manière dépendante du genre à partir des i-vecteurs de dimension 200 (obtenus après la projection par la LDA et la normalisation à **1** de leurs normes euclidiennes) représentant les mêmes ensembles de données utilisés pour entrainer la LDA. De plus, nous avons également estimé ces mêmes paramètres d'une façon indépendants du genre afin de les utiliser par la SAC indépendante du genre (SAC-IG).
- ***Paramètres de la normalisation des scores*** : de la même manière que la matrice de covariance intraclasse et le vecteur moyen, les paramètres de la normalisation des scores

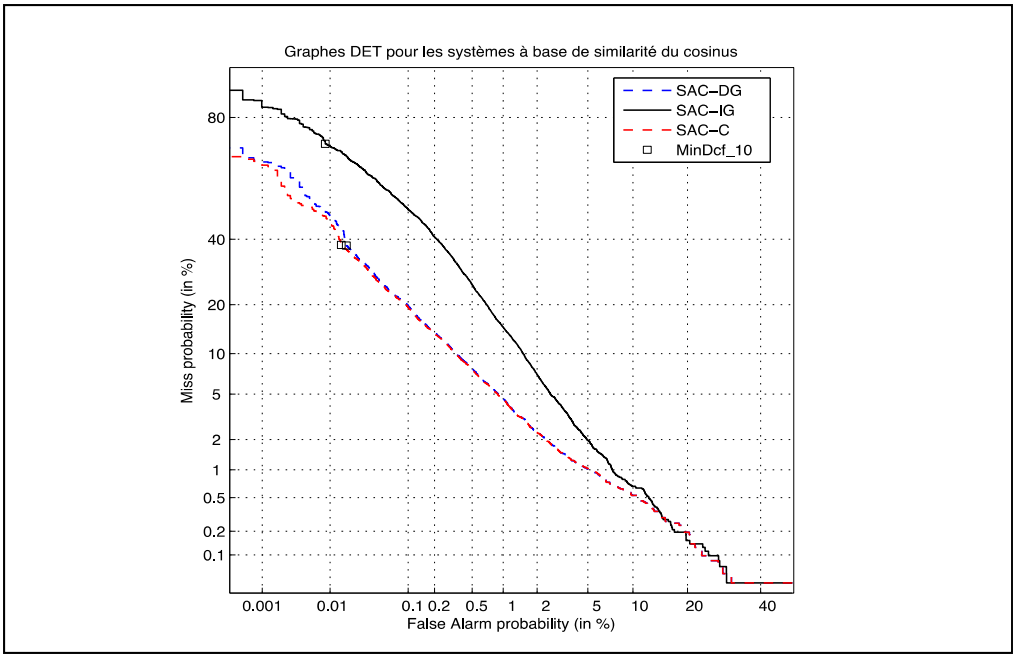

Figure 5.3 Courbes DET des systèmes à base de la SAC testés sur les listes téléphoniques de NIST (det5).

(μ_{imp} et $\mathbf{C}_{imp}$) sont estimés tantôt d'une manière indépendante du genre, tantôt d'une manière dépendante du genre. Ces paramètres sont estimés à partir des mêmes ensembles de données que ceux utilisés pour l'estimation des paramètres de la LDA, de la matrice de covariance intraclasse et du vecteur moyen. En effet, les i-vecteurs à partir desquels nous estimons ces paramètres ont déjà subi la série des transformations précédant la normalisation des scores (voir Figure 5.2), à savoir, la projection via la LDA, la normalisation à **1** des normes euclidiennes, le centrage et la rotation par l'inverse de la matrice de covariance intraclasse.

5.2.3.2 Résultats et discussions

Comme dans les cas des expériences menées sur le modèle génératif PLDA, nous expérimentons d'abord la SAC avec de la parole téléphonique via la liste *det5* de NIST SRE 2010.

Tableau 5.6 Résultats des différents systèmes à base de la SAC testés sur la liste det5 (téléphone/téléphone) de NIST SRE 2010. Les résultats du PLDA-M sont repris à des fins de comparaison.

			EER (%)	MinDCF_08	MinDCF_10
Femmes	SAC-DG	*zt-norm*	**2.62**	0.151	0.551
		s-norm	2.75	**0.136**	0.452
	SAC-IG	*zt-norm*	3.82	0.258	0.698
		s-norm	3.07	0.158	0.517
	SAC-C	*zt-norm*	**2.64**	0.150	0.550
		s-norm	2.73	**0.135**	**0.448**
	PLDA-M	-	2.46	0.124	0.388
Hommes	SAC-DG	*zt-norm*	**1.67**	**0.091**	0.406
		s-norm	1.99	0.102	**0.364**
	SAC-IG	*zt-norm*	2.52	0.164	0.704
		s-norm	2.25	0.133	0.580
	SAC-C	*zt-norm*	**1.67**	**0.091**	0.411
		s-norm	1.98	0.102	**0.368**
	PLDA-M	-	1.81	0.096	0.322

En observant les résultats présentés par le Tableau 5.6, nous pouvons constater que les résultats de la SAC dépendante du genre (SAC-DG) sont constamment semblables à ceux obtenus via la combinaison SAC-C, ce qui prouve son efficacité. De plus, les résultats de la combinaison SAC-C sont clairement meilleurs que ceux du système indépendant du genre à base de la SAC simple (SAC-IG), et ce, tout au long de la courbe DET (voir Figure 5.3). Les résultats des deux méthodes de la normalisation des scores (*zt-norm* et *s-norm*) sont généralement similaires dans le cas de la SAC-DG et de la SAC-C. Enfin, les résultats de la SAC-C sont comparables à ceux obtenus via le modèle génératif du mélange PLDA-M (voir les deux lignes grises du Tableau 5.6).

Les résultats du Tableau 5.7 montrent que la combinaison des SAC proposée dans cette section est également capable de gérer les essais à genre croisé.

Les résultats microphoniques (*det*2) révèlent un comportement similaire à celui des résultats téléphoniques (*det*5) en matière d'équivalence entre les résultats des systèmes dépendants du genre SAC-DG et ceux des systèmes indépendants du genre SAC-C (voir Tableau 5.8). De

plus, nous pouvons percevoir que la méthode *s-norm* de la normalisation des scores a tendance à engendrer de meilleurs résultats en terme d'EER. Tandis que la méthode *zt-norm* a tendance à réduire le minimum de la fonction DCF. Par ailleurs, en observant les résultats dans le Tableau 5.8, il s'avère que le système à base du modèle génératif PLDA-M est plus performant que les systèmes à base de similarité angulaire, et ceci, notamment dans le cas des locutrices.

Tableau 5.7 Résultats des différents systèmes à base de la SAC-C testés sur la liste (téléphone/téléphone) des essais à genre croisé.

		EER (%)	**DCF_08**	**DCF_10**
Femmes	Liste de NIST	2.59	0.149	0.550
	Liste d. g. croisé	**0.95**	**0.084**	**0.507**
Hommes	Liste de NIST	1.66	0.090	0.402
	Liste d. g. croisé	**0.79**	**0.069**	**0.385**

Tableau 5.8 Résultats des différents systèmes à base de la SAC testés sur la liste det2 (téléphone/téléphone) de NIST SRE 2010. Les résultats du PLDA-M sont repris à des fins de comparaison.

			EER (%)	MinDCF_08	MinDCF_10
Femmes	SAC-DG	*zt-norm*	4.53	0.236	0.700
		s-norm	4.73	**0.226**	**0.624**
	SAC-IG	*zt-norm*	7.40	0.443	0.840
		s-norm	4.88	0.256	0.712
	SAC-C	*zt-norm*	**4.51**	0.236	0.695
		s-norm	4.70	**0.225**	**0.622**
	PLDA-M	-	3.87	0.190	0.541
Hommes	SAC-DG	*zt-norm*	**2.23**	0.117	0.493
		s-norm	2.38	**0.114**	**0.414**
	SAC-IG	*zt-norm*	3.80	0.254	0.891
		s-norm	2.69	0.150	0.576
	SAC-C	*zt-norm*	2.25	0.117	0.494
		s-norm	2.39	**0.115**	**0.414**
	PLDA-M	-	**2.03**	**0.097**	**0.365**

Les résultats des expériences menées sur le reste des listes de NIST, *det*1, *det*3 et *det4* (voir Tableau 5.9), reflètent également les mêmes conclusions concernant l'efficacité de la combinaison des similarités (SAC-C) par rapport à la SAC-IG. Encore une fois, le modèle génératif offre de meilleurs résultats par rapport aux systèmes à base de la SAC.

Tableau 5.9 Résultats des différents systèmes à base de la SAC testés sur les listes det1, det3 et det4 de NIST SRE 2010, regroupant les locuteurs hommes et femmes.

			EER (%)	MinDCF_08	MinDCF_10
*det*1	SAC-DG	*zt-norm*	1.89	0.110	0.514
		s-norm	**1.62**	**0.080**	**0.328**
	SAC-IG	*zt-norm*	5.15	0.385	0.856
		s-norm	2.04	0.122	0.552
	SAC-C	*zt-norm*	1.88	0.110	0.515
		s-norm	**1.62**	**0.080**	**0.328**
	PLDA-M	-	1.58	0.070	0.246
*det*3	SAC-DG	*zt-norm*	**3.00**	**0.138**	0.502
		s-norm	3.28	0.146	0.487
	SAC-IG	*zt-norm*	3.90	0.256	0.848
		s-norm	3.25	0.164	0.639
	SAC-C	*zt-norm*	**3.00**	**0.138**	0.491
		s-norm	3.29	0.146	**0.480**
	PLDA-M	-	2.68	0.125	0.397
*det*4	SAC-DG	*zt-norm*	3.68	0.161	0.584
		s-norm	**3.65**	**0.147**	**0.434**
	SAC-IG	*zt-norm*	5.58	0.362	0.889
		s-norm	3.86	0.177	0.608
	SAC-C	*zt-norm*	3.68	0.161	0.585
		s-norm	**3.65**	**0.147**	**0.436**
	PLDA-M	-	2.90	0.129	0.384

Pour conclure, le modèle génératif de mélange PLDA-M proposé est clairement efficace face à toutes les conditions du test. Par ailleurs, la combinaison SAC-C a prouvé son efficacité une fois testée sur les cinq listes d'essais de NIST et également sur la liste d'essais à genre croisé. Les résultats de la parole téléphonique (*det*5) obtenus via la SAC-C sont semblables à ceux obtenus via le modèle génératif PLDA-M. Néanmoins, dans le cas de la parole

microphonique, il s'avère qu'il reste du travail à faire afin d'améliorer les performances des systèmes à base de la SAC.

L'objectif principal de la première partie de cette thèse est de proposer des solutions qui rendent les systèmes de l'état de l'art actuel (c.-à-d. systèmes à base de la PLDA ou bien la SAC dans l'espace des i-vecteurs) indépendant du canal de transmission ainsi que du genre du locuteur. Autrement dit, notre objectif est d'atteindre les performances des systèmes de l'état de l'art déjà existants, sans pour autant tirer profit des informations supplémentaires concernant le type du canal ou le genre du locuteur. Dans le contexte de cet objectif bien précis, nous ne nous sommes pas focalisé sur la comparaison des performances des systèmes qui sont déjà considérés comme les plus performants à l'heure actuelle. De ce fait, seulement les travaux sources sont cités dans le cadre de cette thèse. Or, de nombreux travaux basés sur divers variantes de ces systèmes sont publiés ces dernières années, voir entre autres (Burget, *et al*., 2010)(Dehak, *et al*., 2010)(Bousquet, *et al*., 2012)(Kanagasundaram, *et al*., 2013), etc.

Enfin, les idées présentées dans cette première partie de thèse peuvent être facilement adaptées à d'autres types de classificateurs tels que les SVM, Machines de Boltzmann, les réseaux de neurones, etc.

CHAPITRE 6

L'ALGORITHME DE DÉCALAGE DE LA MOYENNE

6.1 Version de base de l'algorithme du décalage de la moyenne (Mean Shift)

À l'origine, le décalage de la moyenne (*Mean Shift*, MS) est un algorithme itératif de recherche non paramétrique du mode d'une distribution de probabilité. Cependant, cet algorithme peut également servir comme l'unité de base d'un mécanisme de classification/regroupement automatique (*Clustering*). La version originale de l'algorithme de Décalage de la moyenne est présentée dans l'article de Fukunaga (Fukunaga, *et al.*, 1975). Malgré sa première apparition en 1975, le MS est resté longtemps dans l'oubli à l'exception de quelques travaux, entre autres, celui présenté dans (Cheng, 1995) qui vise à généraliser la version originale de Fukunaga. L'algorithme de MS est réapparu ultérieurement en 2002 avec les travaux de Comaniciu en traitement d'image (Comaniciu, *et al.*, 2002). Récemment, les travaux de Stafylakis présentés dans (Stafylakis, *et al.*, 2010)(Stafylakis, *et al.*, 2012) ont montré comment généraliser l'algorithme de MS opérant dans l'espace conventionnel euclidien avec une nouvelle version opérant dans des espaces non euclidiens afin de regrouper des objets complexes autres que les simples points représentés dans l'espace euclidien. Cette nouvelle version d'algorithme de MS a été appliquée au problème de la structuration en tours de parole dans le contexte de la parole diffusée, où les tours de parole des locuteurs ont été caractérisés par des distributions gaussiennes multidimensionnelles. Dans cette section, nous présenterons l'idée intuitive derrière le processus de recherche du mode via le *Décalage de la moyenne* ainsi que les dérivations mathématiques de cet algorithme. Une extension de l'algorithme original de *Décalage de la moyenne* à une nouvelle version à base de la distance angulaire du cosinus (Senoussaoui, *et al.*, 2013 b)(Senoussaoui, *et al.*, 2014) sera également détaillée. En outre, nous présenterons deux variantes des mécanismes exploitant cet algorithme à des fins de regroupement des données non étiquetées. Finalement, il est important de souligner que la représentation de la parole par des i-vecteurs (Dehak, *et al.*, 2011b) sera adoptée dans toutes les expériences menées dans le but de tester les différentes variantes d'algorithmes de *Décalage de la moyenne*.

6.1.1 Idée intuitive

Le *Décalage de la moyenne* est un membre de la famille des estimateurs non paramétriques de densités de probabilité, ou ce qu'on dénomme aussi, les estimateurs de densités à base de noyaux (les méthodes de Parzen-Rozenblatt). À l'inverse des estimateurs de densités (KDE) qui cherchent à estimer la valeur de la densité de probabilité autour d'une observation donnée, le *Décalage de la moyenne* cherche simplement à atteindre le mode de cette densité et non pas sa valeur (voir Figure 6.1). Pour ce faire, le *Décalage de la moyenne* se base sur le calcul du gradient de la fonction de densité de probabilité (Fukunaga, *et al.*, 1975)(Comaniciu, *et al.*, 2002).

L'idée intuitive de *Décalage de la moyenne* est tout à fait naturelle et simple. À partir d'un vecteur quelconque $\mathbf{x}_i$ appartenant à un ensemble $S = \{\mathbf{x}_i\}_{i=1..n}$ de n données non étiquetées (qui sont en réalité des i-vecteurs dans notre cas), nous pouvons atteindre via le processus itératif décrit dans l'**Algorithme 1** un point stationnaire dénommé le mode de la densité. Notons que l'**Algorithme 1** présente la version originale de *Décalage de la moyenne*.

Algorithme 1 *Décalage de la moyenne – Idée intuitive*

- $i = 1$, $\mathbf{y}_i = \mathbf{x}_i$
- Centrer une fenêtre noyau autour de $\mathbf{y}_i$ *//Initialisation*

Répéter

- $\mu_h(\mathbf{y}_i)$ *//estimer le vecteur moyen des données qui se trouvent à l'intérieur de la fenêtre (c.-à-d. les points voisins de* $\mathbf{y}_i$ *en terme de la distance euclidienne)*
- $y_{i+1} = \mu_h(\mathbf{y}_i)$
- déplacer la fenêtre de $\mathbf{y}_i$ à $\mathbf{y}_{i+1}$
- $i = i+1$

Jusqu'à : Stabilisation *//un mode a été trouvé*

La preuve mathématique de convergence de la séquence $\{\mathbf{y}_i\}_{i=1,2...}$ des positions successives se trouve dans les articles (Fukunaga, *et al.*, 1975)(Comaniciu, *et al.*, 2002).

6.1.2 Développement mathématique

L'estimation de la fonction de densité de probabilité en se basant seulement sur une faible quantité de données autour d'une observation donnée est un problème largement connu dans le domaine de la reconnaissance de formes. Étant donné l'ensemble $S = \{\mathbf{x}_i\}_{i=1..n}$ de n observations non étiquetées, la forme standard d'un estimateur à base de noyau de la fonction de densité de probabilité $\hat{f}(\mathbf{x})$ autour d'une observation $\mathbf{x}$ est donnée par la formule suivante :

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \tag{6.1}$$

où $k(\mathbf{x})$ est un noyau symétrique de largeur radiale h, ou ce qu'on dénomme habituellement la bande passante. Son rôle est le lissage de la fonction de densité estimée. Effectivement, si on ignore le choix du type de noyau, le scalaire h reste le seul hyper-paramètre nécessitant une estimation sur un ensemble indépendant de développement. Dans le but d'assurer quelques propriétés comme l'estimation asymptotique non biaisée et la consistance, le noyau de largeur h doit satisfaire quelques conditions qui sont bien détaillées dans (Fukunaga, *et al.*, 1975).

Tel que préalablement évoqué, la recherche du mode de la fonction de densité $f(\mathbf{x})$nécessite l'évaluation du gradient de cette fonction. En fait, l'estimation du gradient de la fonction de densité$\hat{\nabla} f(\mathbf{x})$est approximée par le gradient de la fonction de densité estimée $\nabla\hat{f}(\mathbf{x})$comme suit :

$$\begin{aligned}\hat{\nabla} f(\mathbf{x}) \equiv \nabla\hat{f}(\mathbf{x}) &= \frac{1}{nh^d}\sum_{i=1}^{n}\nabla k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \\ &= \frac{2}{nh^{d+2}}\sum_{i=1}^{n}(\mathbf{x}-\mathbf{x}_i)k'\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right).\end{aligned} \tag{6.2}$$

Prenons le simple exemple du noyau d'Epanechnikov dont le profil est donné par l'équation suivante :

$$k(\mathbf{x}) = \begin{cases} 1 - \|\mathbf{x}\|^2 & \|\mathbf{x}\| \le 1 \\ 0 & \|\mathbf{x}\| > 1 \end{cases}. \tag{6.3}$$

le noyau uniforme donné par :

$$g(\mathbf{x}) = \begin{cases} 1 & \|\mathbf{x}\| \le 1 \\ 0 & \|\mathbf{x}\| > 1 \end{cases} \tag{6.4}$$

satisfait la condition suivante :

$$k'(\mathbf{x}) = -c\, g(\mathbf{x}) \tag{6.5}$$

où c est une constante de normalisation quelconque.

Dans le cas du noyau d'Epanechnikov, le gradient est donné par :

$$\begin{aligned} \nabla \hat{f}(\mathbf{x}) &= \frac{1}{nh^{d+2}} \sum_{i=1}^{n} \nabla k\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \\ &= \frac{2}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_i) k'\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \\ &= \frac{2}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}) g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \\ &= \frac{2}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x} \right]. \end{aligned} \tag{6.6}$$

La partie entre crochets la plus à gauche de l'équation 6.6 donnée par

$$\mathbf{m}_h(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x} \tag{6.7}$$

est nommée le vecteur de *Mean Shift* $\mathbf{m}_h(\mathbf{x})$.

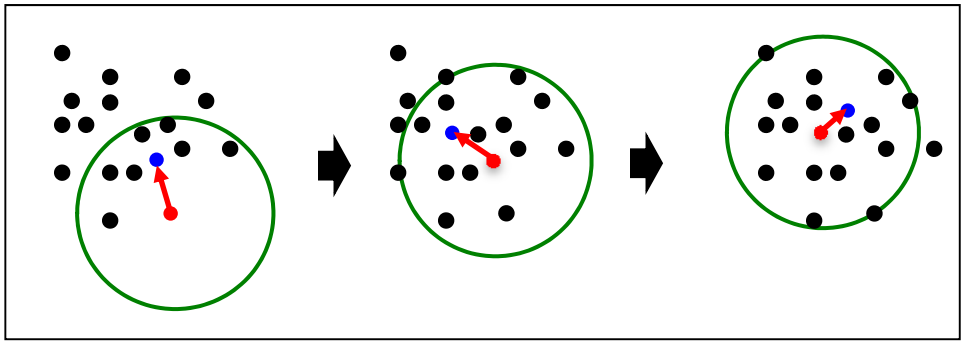

Figure 6.1 Évolution du processus de recherche du mode d'une distribution de probabilité via l'algorithme de *Décalage de la moyenne*. Le cercle vert représente le noyau centré sur le point d'intérêt (point rouge), la flèche rouge représente le vecteur de *Mean Shift* et finalement le point bleu représente la moyenne des points qui se trouvent dans le cercle de noyau.

En observant le vecteur de *Mean Shift* $\mathbf{m}_h(\mathbf{x})$ donné par l'équation (6.7), nous constatons qu'il n'est en fait que la différence entre deux vecteurs. Le premier représente la position courante ou bien la position d'intérêt (c.-à-d. le point $\mathbf{x}$ autour duquel nous voulons estimer le mode de la distribution de probabilité). Le deuxième représente la prochaine position donnée par un vecteur moyen pondéré de l'ensemble des données (voir Figure 6.1). À vrai dire, les poids de pondération sont les valeurs binaires (c.-à-d. 0 ou 1) émises par le noyau uniforme.

Pour des raisons de simplicité, nous allons noter le noyau uniforme de paramètre h par $g(\mathbf{x},\mathbf{x}_i,h)$ de sorte que :

$$g(\mathbf{x},\mathbf{x}_i,h)=\begin{cases}1 & \|\mathbf{x}-\mathbf{x}_i\|^2 \le h^2 \\ 0 & \|\mathbf{x}-\mathbf{x}_i\|^2 > h^2\end{cases} . \tag{6.8}$$

Le rôle du noyau uniforme est de sélectionner un sous-ensemble S_x de données qui se trouvent à l'intérieur du cercle de noyau centré sur $\mathbf{x}$. Autrement dit, S_x contient des observations dont les distances euclidiennes par rapport au point d'intérêt $\mathbf{x}$ sont inférieures ou égales à la largeur de la bande passante h, comme suit :

$$S_h(\mathbf{x}) \equiv \{\mathbf{x}_i : \|\mathbf{x}_i-\mathbf{x}\| \le h\} . \tag{6.9}$$

Par conséquent, nous pouvons reformuler le vecteur de *Mean Shift* $\mathbf{m}_h(\mathbf{x})$ de la manière suivante :

$$\mathbf{m}_h(\mathbf{x}) = \mu_{S_h(\mathbf{x})} - \mathbf{x} \tag{6.10}$$

où $\mu_{S_h(\mathbf{x})}$ est le vecteur moyen des observations de l'ensemble S_h(x) (voir Figure 6.1).

L'exécution itérative du calcul de la moyenne des échantillons limités par un noyau suivi par le décalage de ce noyau vers la moyenne calculée (ce qui produit la séquence $\{\mathbf{y}_i\}_{i=1,2...}$ des positions successives du noyau présentées par les points rouges dans la Figure 6.1) aboutit à un point stationnaire représentant le mode de la distribution des données.

6.2 Algorithme de Décalage de la moyenne à base de distance angulaire

6.2.1 Motivations

Le remplacement de la distance euclidienne conventionnelle dans l'algorithme de base de *Décalage de la moyenne* par la distance angulaire du cosinus est fortement motivé par le succès de cette distance dans plusieurs domaines connexes, tels que la vérification du locuteur (Dehak, *et al.*, 2010)(Dehak, *et al.*, 2011b)(Senoussaoui, *et al.*, 2013a), l'identification de la langue (Dehak, *et al.*, 2011c) et même la structuration en tours de parole (Shum, *et al.*, 2011)(Shum, *et al.*, 2012)(Shum, *et al.*, 2013). Notons également que dans l'article (Tang, *et al.*, 2012), les chercheurs ont montré que la distance du cosinus offre une meilleure métrique que la distance euclidienne dans l'espace des supervecteurs.

De plus, la nature gaussienne standard qui caractérise la distribution des i-vecteurs (qui sont utilisés pour représenter les segments vocaux dans tous nos travaux) constitue un autre argument justifiant l'adoption de la distance du cosinus au lieu de la distance euclidienne. Supposons que nous sommes en possession d'une paire des i-vecteurs et nous voulons tester l'hypothèse H_0 stipulant qu'ils appartiennent au même groupe (locuteur) contre l'hypothèse H_1 stipulant qu'ils appartiennent à différents groupes. Étant donné que la masse

de la population des i-vecteurs est principalement concentrée dans le voisinage de l'origine d'espace, les locuteurs dans cette région sont plus susceptibles d'être confondus les uns avec les autres. Donc, dans le cas d'une paire des i-vecteurs proches de l'origine, hypothèse H_0 ne sera admise que si les i-vecteurs sont relativement rapprochés. En outre, si ces i-vecteurs sont loin de l'origine, ils peuvent être relativement éloignés l'un de l'autre tout en conservant la validité de la même hypothèse H_0. Par conséquent, il est essentiel d'intégrer cette connaissance *a priori* concernant la distribution des locuteurs dans l'algorithme de *Décalage de la moyenne*. Ceci peut être réalisé par l'utilisation de la distance euclidienne et une bande passante h variable, qui augmente avec l'éloignement de l'origine, ou bien par l'utilisation d'une bande passante fixe et la distance du cosinus. Évidemment, la deuxième approche est préférable.

6.2.2 Développement mathématique

La distance angulaire du cosinus $\mathcal{D}$ entre deux i-vecteurs $\mathbf{x}_1$ et $\mathbf{x}_2$ est donnée par :

$$\mathcal{D}(\mathbf{x}_1,\mathbf{x}_2) = 1 - \left(\frac{\mathbf{x}_1' \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_1\|} \right) \tag{6.11}$$

L'algorithme original de *Décalage de la moyenne* est basé sur un noyau uniforme. Une fois la fenêtre de ce noyau est centrée sur un point d'intérêt $\mathbf{x}$, tout point dont la distance euclidienne par rapport au point $\mathbf{x}$ est inférieure ou égale à la largeur de la bande h, est considéré comme un point se situant à l'intérieur de cette fenêtre (voir équation 6.9). Afin d'aboutir à la nouvelle version de l'algorithme de *Décalage de la moyenne* (MS à base de la distance euclidienne), une seule modification (Senoussaoui, *et al.*, 2013b)(Senoussaoui, *et al.*, 2014) doit être introduite au niveau de l'équation (6.9) comme suit :

$$S_h(\mathbf{x}) \equiv \left\{ \mathbf{x}_i : \mathcal{D}(\mathbf{x},\mathbf{x}_i) \leq h \right\} \tag{6.12}$$

où $\mathcal{D}(\mathbf{x}, \mathbf{x}_i)$ est la distance angulaire du cosinus donnée par l'équation (6.11). Cette modification mène naturellement à modifier la reformulation du noyau uniforme donnée par l'équation (6.8) comme suit :

$$g(\mathbf{x}, \mathbf{x}_i, h) = \begin{cases} 1 & D(\mathbf{x}, \mathbf{x}_i) \le h \\ 0 & D(\mathbf{x}, \mathbf{x}_i) > h \end{cases}. \quad (6.13)$$

6.3 Algorithme de Décalage de la moyenne pour le regroupement des données non étiquetées

Comme évoqué précédemment, l'algorithme de *Décalage de la moyenne* est principalement conçu pour des fins de recherche non paramétrique du mode d'une distribution inconnue. Or, le MS peut être aussi utilisé au sein d'un mécanisme de regroupement (classification) automatique des données non étiquetées formant un nombre inconnu de classes (Fukunaga, *et al.*, 1975). Il est aussi possible d'utiliser cet algorithme pour résoudre d'autres problèmes tels que la segmentation des images et le suivi des objets (Comaniciu, *et al.*, 2002). Dans notre recherche, nous nous intéressons particulièrement à la tâche du regroupement qui peut être accomplie au moyen d'un mécanisme exploitant l'algorithme de MS. Ainsi, nous présenterons dans les sous-sections qui suivent deux stratégies de regroupement automatique à base de l'algorithme de MS, à savoir, la *stratégie totale* de regroupement et la *stratégie sélective* de regroupement.

6.3.1 Stratégie totale de regroupement (STR)

Afin de regrouper automatiquement un ensemble de données non étiquetées, on peut appliquer la procédure itérative de *Décalage de la moyenne* à chaque observation de cet ensemble de données. D'une manière générale, certains processus de MS convergent au même mode de densité. Par conséquent, le nombre des modes uniques de densité détectés sur l'ensemble des données (après une opération d'élagage) représente le nombre des classes détectées. Évidemment, les points, dont les processus MS associés convergent au même mode, sont assignés à la même classe (l'ensemble de ces points est nommé le bassin d'attraction « *basin of attraction* » de mode associé). Dans ce travail, nous dénommons cette approche la *stratégie totale* de regroupement (STR).

6.3.2 Stratégie sélective de regroupement (SSR)

Contrairement à la *stratégie totale* de regroupement (STR), nous pouvons adopter cette deuxième stratégie, dénommée la *stratégie sélective* de regroupement (SSR), afin d'appliquer le processus MS seulement à un sous-ensemble de données à regrouper. L'idée derrière cette stratégie est de conserver, tout au long de l'évolution d'un processus MS, une trace du nombre de visites de chaque point de l'ensemble des données. Après la convergence du premier processus MS, tous les points qui ont été visités durant ce processus seront affectés à cette première classe détectée. De la même manière, nous amorçons un deuxième processus à partir de l'un des points non visités au cours du premier processus. Nous continuons d'exécuter ces processus l'un après l'autre jusqu'à ce que nous n'ayons aucun point non visité de notre ensemble de départ. Il est fort probable que certains points puissent être assignés par cette stratégie à plus d'une seule classe, ainsi, une opération de vote majoritaire basé sur le nombre des visites conservé pour chaque processus MS est nécessaire afin de concilier ces conflits.

Il est important de souligner que la complexité du calcul dans le cas de la *stratégie totale* (STR) dépend du nombre total des points de l'ensemble des données à regrouper. Par ailleurs, cette complexité, dans le cas de la *stratégie sélective* (SSR), ne dépend que du nombre des classes détectées. Finalement, une implémentation MATLAB de la *stratégie sélective* peut être trouvée en ligne sur ce lien[13].

[13] http://www.mathworks.com/matlabcentral/fileexchange/authors/22444

CHAPITRE 7

REGROUPEMENT EN LOCUTEURS

7.1 Regroupement en locuteurs

Le regroupement automatique (*Clustering*) des données non étiquetées est une tâche fondamentale pour plusieurs disciplines, notamment pour celles qui se basent sur les concepts de la reconnaissance des formes et de l'apprentissage-machine. L'objectif ultime du regroupement automatique d'un ensemble de données est de relier les observations similaires selon une métrique donnée afin de déterminer les regroupements intrinsèques de cet ensemble. Autrement dit, on peut reformuler cet objectif de la manière suivante : étant donné un ensemble de données non étiquetées, où chaque observation a été générée par une classe unique (groupe) et chaque classe a une ou plusieurs observations. Le processus du regroupement tente de : i) assigner toutes les observations originaires d'une même classe au même groupe, ii) avoir des groupes dont chacun contient les observations originaires de la même classe.

Dans le cadre du traitement de la parole, on tente souvent d'étiqueter les segments vocaux selon leurs contenus linguistiques (tels que phonèmes, mots, expressions, etc.) ou bien selon l'état de leurs locuteurs émetteurs (émotion, genre, santé, âge, etc.) ou bien même parfois selon les identités des locuteurs. Lorsque l'étiquette correspond à l'identité du locuteur, cette tâche est appelée regroupement en locuteurs (*Speaker Clustering*). Ainsi, le regroupement en locuteurs d'un ensemble de segments vocaux non étiquetés est défini par le fait d'assigner à chaque segment une étiquette (voir Figure 1.5) relative à l'identité du locuteur présumé (Kotti, *et al.*, 2008)(Van Leeuwen, 2010)(Senoussaoui, *et al.*, 2013b). À vrai dire, cette identité peut être vraie comme elle peut être arbitraire.

Le regroupement en locuteurs est une discipline importante pour de nombreux domaines du traitement de la parole. Citons à guise d'exemple, l'adaptation automatique des systèmes de la reconnaissance vocale indépendante du locuteur à la parole d'un locuteur donné afin

d'améliorer les performances de ces systèmes. Le regroupement en locuteurs peut également servir à augmenter la capacité d'archivage et de stockage des fichiers audio dans les grandes bases de données multimédias et y accélérer la recherche.

7.2 Méthodologie

D'une manière générale, la tâche du regroupement automatique s'achève par des algorithmes dotés d'un mécanisme d'agrégation à base d'une métrique qui mesure la similarité/distance entre les observations dans un espace des caractéristiques. En effet, le choix de l'espace des caractéristiques, du mécanisme d'agrégation et de la métrique, influence directement les performances du système de regroupement. À titre d'exemple, l'auteur Van Leeuwen (Van Leeuwen, 2010) a adopté l'espace des supervecteurs comme un espace des caractéristiques, une méthode à base du regroupement hiérarchique comme un mécanisme d'agrégation et le score d'un système de vérification à base de machines à vaste marge (SVM) comme une métrique mesurant la ressemblance entre les observations.

7.2.1 Représentation du signal vocal

Nous commençons d'abord par le choix de l'espace de représentation (espace des caractéristiques). Dans le contexte actuel de la recherche dans le domaine de la reconnaissance du locuteur et spécialement après la découverte des i-vecteurs (Dehak, *et al.*, 2011b), la question du choix de l'espace de représentation est devenue triviale. Ainsi, l'espace des i-vecteurs est adopté afin de représenter chaque segment vocal par un vecteur de faible dimension.

7.2.2 Décalage de la moyenne à base de la distance angulaire du cosinus

Traditionnellement, les méthodes à base de regroupement hiérarchique (*Hierarchical Agglomerative Clustering*, HAC) sont les plus répandues dans l'implémentation des systèmes du regroupement en locuteurs (Kotti, *et al.*, 2008)(Van Leeuwen, 2010). Récemment, l'algorithme de décalage de la moyenne (MS) a commencé de gagner du terrain dans le

domaine du regroupement en locuteurs (Stafylakis, *et al.*, 2010)(Stafylakis, *et al.*, 2012)(Senoussaoui, *et al.*, 2013b)(Senoussaoui, *et al.*, 2014). De par sa nature non paramétrique, l'algorithme MS a l'avantage principal de ne requérir aucune forte hypothèse sur la forme de la distribution des groupes de données. De plus, la procédure itérative du MS s'appuie sur une preuve mathématique de convergence vers le mode de la distribution des données. Finalement, le seul hyper-paramètre nécessitant une estimation sur un ensemble de développement indépendant est le simple scalaire *h* qui définit la largeur de la bande (voir Section 6.1.2).

Dans ce travail, la *stratégie totale* de regroupement (STR) exploitant le processus MS à base de la distance angulaire du cosinus (voir Section 6.3.1) sera notre mécanisme d'agrégation et bien évidemment, la distance angulaire du cosinus sera la métrique adoptée.

7.3 Expérimentation

D'abord, il est important de souligner que l'intérêt accordé à la tâche du regroupement en locuteurs dans cette thèse tend davantage vers la validation de l'efficacité de la nouvelle version de l'algorithme de *Décalage de la moyenne* (c.-à.-d le MS à base de la distance du cosinus) face à la tâche du regroupement en locuteurs. Par ailleurs, cette version de MS sera encore décortiquée dans le contexte de la structuration en tours de parole (*Speaker Diarization*) au Chapitre 8. La principale différence entre la tâche du regroupement en locuteurs des grandes bases de données et celle de la structuration en tours de parole d'un flux audio est le fait que, dans la première, les différents segments à regrouper sont enregistrés sur divers fichiers (c.-à-d. enregistrés durant différentes sessions). Cependant, dans la deuxième, les segments sont enregistrés sur un seul fichier audio. De ce fait, les effets du canal affectant les segments vocaux dans les deux cas ne sont pas semblables. Dans le cadre de ce chapitre, nous nous focalisons seulement sur la compensation des effets nuisibles du canal dans le cas du regroupement en locuteurs des grands corpus de données.

7.3.1 Compensation des effets du canal

Comme évoqué ci-dessus, les segments à regrouper sont enregistrés sur différentes sessions dans le cas du regroupement en locuteurs. Ceci signifie que la variabilité intralocuteur (exprimée principalement par la variabilité intersessions) est à la fois considérable et nuisible. Ainsi, il est inévitable de la compenser au niveau des i-vecteurs, et ce, avant d'exécuter l'algorithme de MS.

Étant donné que l'espace des caractéristiques considéré dans ce travail est celui des i-vecteurs et que la métrique adoptée afin de mesurer la distance entre les observations est la distance angulaire du cosinus. La meilleure stratégie de compensation des effets indésirables du canal est celle utilisée dans le cas d'un système de vérification du locuteur à base de la similarité angulaire du cosinus. Cette stratégie est en fait basée sur la projection des i-vecteurs dans l'espace de dimension réduite formé par l'Analyse discriminante linéaire (LDA) suivie par une rotation via la décomposition de l'inverse de la matrice de covariance intraclasse (voir Section 3.2.1).

7.3.2 Détails d'implémentation

7.3.2.1 Corpus de données du test

Afin de tester l'efficacité de la version proposée de *Décalage de la moyenne* (c.-à-d. MS à base de la distance du cosinus), nous utiliserons les données téléphoniques du test fournies par NIST durant son évaluation de la reconnaissance du locuteur en 2008. Au total, cette base contient 3090 enregistrements téléphoniques de $N = 1270$ locuteurs des deux genres. Chaque enregistrement contient environ 2.5 minutes de parole d'un seul locuteur et évidemment chaque locuteur a au moins un enregistrement.

7.3.2.2 Procédure expérimentale

Cette série d'expériences est menée dans le but d'évaluer et de comparer les performances de la nouvelle version d'algorithme de *Décalage de la moyenne* avec la version de base. D'abord, en utilisant la nouvelle version de MS (c.-à-d. le MS à base de la distance du cosinus), nous avons effectué le regroupement des données de test à plusieurs reprises tout en changeant la largeur de la bande h de 0,1 à 0,99. De manière analogue, nous avons refait la même procédure avec la version originale de MS (c.-à-d. le MS à base de la distance euclidienne) en changeant, cette fois-ci, la largeur de la bande h de 10 à 35.

7.3.2.3 Extraction et normalisation des i-vecteurs

La configuration des composantes d'extracteur des i-vecteurs (c.-à-d. le modèle du monde UBM et l'extracteur des i-vecteur) ainsi que celle des méthodes de normalisation (c.-à-d. la LDA et la WCCN) sont exactement pareilles à celles utilisées dans les expériences du Chapitre 5 (voir Section 5.1.3.1 et Section 5.2.3.1). Pour être précis, tous ces paramètres sont estimés d'une manière indépendante du genre du locuteur à partir des données (locuteurs hommes et femmes regroupés) du développement de NIST. Enfin, nous rappelons que la dimension d'origine d'espace des i-vecteurs est de 800. Elle est, toutefois, réduite à 200 après la projection dans l'espace de la LDA.

7.3.2.4 Métriques d'évaluation des performances

Nous mesurons les performances de nos systèmes de regroupement en locuteurs par les deux métriques à base d'impuretés détaillées dans la Section 1.3.2, soit, l'impureté de classe (I_c) et l'impureté du locuteur (I_s) (Van Leeuwen, 2010). Étant donné la relation inverse reliant les deux impuretés, le point d'égalité ($I_c = I_s$) qui ressemble au EER (point d'égalité des deux taux d'erreurs dans le cas de la vérification du locuteur) représente le point optimal des impuretés. Le nombre de groupes (locuteurs dans notre cas) détectés N_c est un autre indice des performances du système du regroupement en locuteurs. Plus le nombre N_c,

correspondant au point d'égalité des impuretés, est proche du nombre réel N des locuteurs présents dans la base de données à regrouper, plus on juge que le comportement du système de regroupement est meilleur.

7.3.3 Résultats et discutions

Dans le Tableau 7.1, nous présentons uniquement les résultats les plus intéressants à partir desquels nous pouvons observer les points d'égalités des impuretés (les cellules en caractères gras) ainsi que leurs nombres de groupes estimés N_c (les lignes soulignées en gris).

Tableau 7.1 Résultats des deux systèmes du regroupement en locuteurs tels que mesurés par les deux types d'impuretés et le nombre des groupes détectés. Le nombre réel des classes (locuteurs) $N = 1270$.

Distance *euclidienne*				***Distance du cosinus***			
h	I_c	I_s	N_c	h	I_c	I_s	N_c
23.8	0.238	0.085	1297	0.46	0.207	0.060	1161
23.7	0.215	0.090	1337	0.45	0.168	0.065	1225
23.6	0.199	0.096	1368	0.44	0.137	0.071	1286
23.5	0.178	0.101	1397	0.43	0.109	0.080	1352
23.4	0.153	0.108	1438	**0.42**	**0.089**	**0.092**	**1414**
23.3	0.142	0.113	1461	0.41	0.069	0.106	1471
23.2	**0.126**	**0.120**	**1496**	0.40	0.056	0.120	1537
23.1	0.105	0.128	1533	0.39	0.047	0.135	1602

L'observation des résultats présentés par le Tableau 7.1 nous permet d'en tirer quelques conclusions. Tout d'abord, il est clair que les deux systèmes de regroupement atteignent leurs points d'égalités des impuretés (0,123 pour la version de MS à base de la distance euclidienne contre 0,09 pour la version de MS à base de la distance angulaire du cosinus) après une surestimation du nombre des groupes estimés N_c (1496 pour la version de MS à base de la distance euclidienne contre 1414 pour la version de MS à base de la distance angulaire du cosinus). Or, la surestimation du nombre des classes dans le cas de la version de

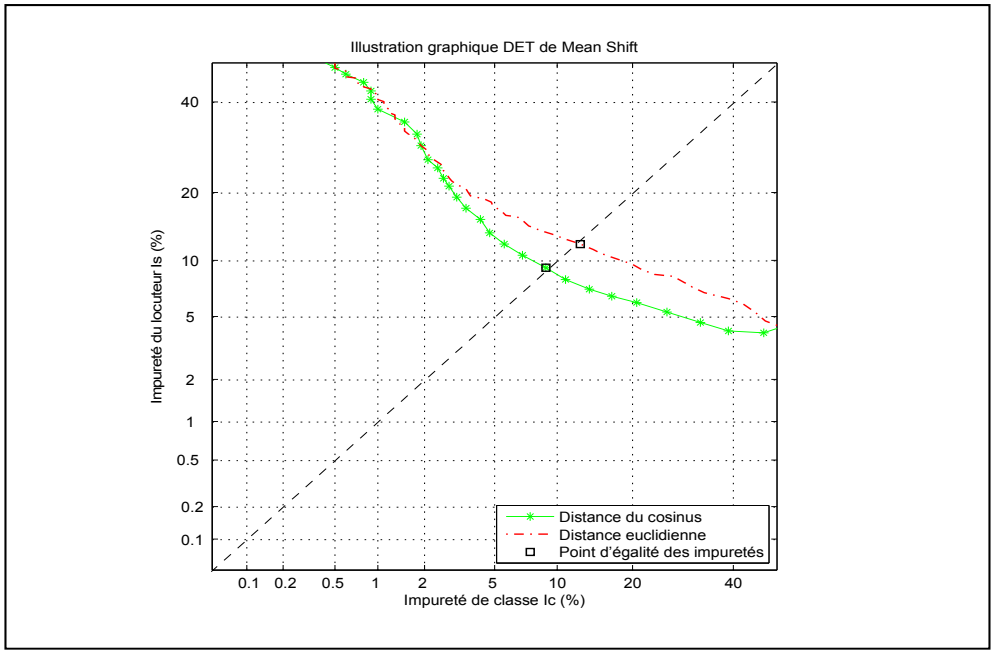

Figure 7.1 Illustration graphique DET des performances de l'algorithme de Mean Shift à base de la distance euclidienne et de celui à base de la distance angulaire du cosinus testés sur la tâche du regroupement en locuteurs de l'ensemble des données SRE 2008 de NIST.

MS à base de la distance angulaire du cosinus est légèrement moins notable que celle de la version originale de MS. Le N_c estimé par la version de MS à base de la distance angulaire du cosinus est ainsi plus proche du nombre réel N que celui estimé par la version originale. Il est important de souligner que dans le cadre général du regroupement en locuteurs, la surestimation du nombre des classes est souvent préférable à sa sous-estimation.

En termes de point d'égalité des impuretés, il est clair que les performances du système à base de la distance angulaire du cosinus sont meilleures que celles du système à base de la distance euclidienne (0,123 pour la version de MS à base de la distance euclidienne contre 0,09 pour la version de MS à base de la distance angulaire du cosinus). D'une manière analogue à celle présentée dans (Van Leeuwen, 2010), si on considère que le point d'égalité des impuretés est un pourcentage semblable à EER de NIST (Equal Error Rate, EER) utilisé dans le cas des systèmes de la vérification du locuteur, nous obtenons une amélioration relative des performances de 11 % (c.-à-d. de 12,3 %, dans le cas euclidien, à 9 %, dans le cas du cosinus).

Malheureusement, nous ne sommes pas en mesure d'effectuer une comparaison réelle avec les résultats rapportés dans l'article (Van Leeuwen, 2010), car le corpus du test considéré dans cet article est celui de NIST SRE 2006. En effet, pour la communauté de la reconnaissance du locuteur, il est bien répandu que la tâche de vérification du locuteur est largement plus facile à partir de l'ensemble des données de NIST SRE 2006 qu'à partir de celui de NIST SRE 2008. Néanmoins, l'auteur (Van Leeuwen, 2010) a obtenu un point d'égalité des impuretés égal à 13,9 % comparativement au nôtre qui est égal à 9 %.

Enfin, les courbes *DET* des deux systèmes (représentés par la Figure 7.1) montrent clairement que les performances du système à base de la distance angulaire du cosinus sont meilleures que celles du système à base de la distance euclidienne, et ce, pour la quasi-totalité des courbes, et plus particulièrement dans la zone de faible impureté de locuteur.

CHAPITRE 8

STRUCTURATION EN TOURS DE PAROLE

8.1 Structuration en tours de parole

La tâche de la structuration automatique en tours de parole consiste à fractionner un flux audio c.-à-d. un fichier audio contenant un discours de plusieurs locuteurs en des zones homogènes qui correspondent à la parole de chacun des locuteurs participants. En effet, on veut répondre aux questions « qui parle ? quand ? » lorsque la structuration automatique en tours de parole est combinée avec la reconnaissance automatique du locuteur.

De la façon dont le problème est habituellement formulé, la structuration en tours de parole nécessite l'achèvement de deux sous-tâches principales, à savoir, la s*egmentation* et le *regroupement*. La segmentation est la phase du repérage des frontières entre les tours de parole afin de produire un ensemble de segments non chevauchés du flux audio. Effectivement, à ce stade, les segments produits par cette phase sont considérés comme étant indépendants les uns des autres. La phase de regroupement, quant à elle, vise à lier, selon une mesure donnée, les segments non étiquetés, issus de l'étape de segmentation, afin de déterminer les masses intrinsèques dans l'ensemble de ces segments.

Certainement, l'absence d'information *a priori* concernant le nombre de locuteurs dans le flux amplifie considérablement la difficulté de la tâche de structuration. En outre, la courte durée des tours de parole dans le cas du dialogue téléphonique (habituellement une seconde) rend la tâche de la représentation de ces tours de parole dans l'espace des caractéristiques assez difficile. La variabilité de la durée des fichiers audio à segmenter peut également jouer le rôle d'un autre facteur dégradant des performances.

8.2 Méthodologie

La sélection du modèle basée sur le critère d'information bayésien (*Bayesian Information Criterion*, BIC) est la méthode la plus populaire pour réaliser la *segmentation* des flux audio

(Schwarz, 1978)(Chen, *et al.*, 1998). De plus, le critère BIC sert aussi à estimer le nombre de locuteurs dans un segment vocal. Dans le cadre de la structuration en tours de parole, le BIC n'est guère la seule méthode de nature bayésienne utilisée, d'autres méthodes bayésiennes ont été proposées récemment (Valente, 2005)(Kenny, *et al.*, 2010b). Tout comme le regroupement en locuteurs de grands corpora de données, les méthodes à base du regroupement hiérarchique (*Hierarchical Agglomerative Clustering*, HAC) sont couramment adoptées afin de faire face au problème de la structuration en tours de parole. D'autres méthodes, y compris des approches hybrides, continuent d'être développées pour remédier à ce problème (Kotti, *et al.*, 2008)(Kenny, *et al.*, 2010b)(Shum, *et al.*, 2013).

Dans l'article (Shum, *et al.*, 2011), les auteurs proposent un système de structuration en tours de parole où les i-vecteurs sont utilisés pour représenter les tours de parole et l'algorithme des k-moyennes à base de la distance du cosinus pour associer ces tours de parole avec les identités individuelles des locuteurs impliqués. Testées sur des conversations téléphoniques à deux locuteurs, les performances de cette approche ont dépassé largement celles d'un système basé sur le critère BIC et le mécanisme de regroupement hiérarchique (HAC). Toutefois, afin que l'algorithme des k-moyennes fonctionne, le nombre de locuteurs impliqués dans une conversation donnée doit être connu *a priori*. Ainsi, l'extension de cette approche au problème général de la structuration en tours de parole, où le nombre de locuteurs participant doit être déterminé, n'est guère évidente. Pour ce faire, les auteurs de (Shum, *et al.*, 2011) ont présenté une simple heuristique. Notre principale contribution est de démontrer comment l'utilisation de *Décalage de la moyenne* au lieu de k-moyennes permet de traiter efficacement ce problème.

Dans le cadre de ce travail, nous nous focalisons sur la sous-tâche du *regroupement* plutôt que celle de la *segmentation*. Nous avons également choisi la parole téléphonique afin d'expérimenter nos algorithmes du regroupement à cause du plus grand défi de cette tâche. Tel que évoqué ci-dessus, le discours téléphonique est principalement caractérisé par la courte durée de ces tours de parole. Usuellement, les locuteurs parlent environ une seconde avant de faire une pause ou de passer la parole à leurs interlocuteurs (Kenny, *et al.*, 2010b).

Cette spécificité rend plus complexe la tâche de la représentation de ces segments de très courtes durées dans un espace tel que celui des i-vecteurs.

Dans le but d'entreprendre la question de la structuration d'un flux téléphonique en tours de parole, nous représentons d'abord chaque segment (c.-à-d. un tour de parole) par un i-vecteur. Ensuite, ces i-vecteurs seront regroupés via l'algorithme de *Décalage de la moyenne* à base de la distance du cosinus (voir Section 6.2 et Figure 8.1). Dans le cadre de cette recherche, des méthodes de normalisation des i-vecteurs en vue d'atténuer les effets indésirables du canal seront discutées. Du plus, nous proposons deux solutions destinées respectivement à remédier au problème de l'unique bande passante et à celui de la surestimation du nombre de locuteurs. En effet, ces problèmes dégradent souvent les performances de regroupement à base de l'algorithme MS. Finalement, les comportements des deux mécanismes de regroupement (voir Section 6.3) associés à l'algorithme de *Décalage de la moyenne* à base de la distance du cosinus, à savoir, la *stratégie totale* du regroupement (STR) et la *stratégie sélective* du regroupement (SSR), seront également examinés dans une étude expérimentale comparative. En effet, nos résultats seront comparés avec ceux de l'état de l'art en la matière.

8.2.1 Segmentation initiale en tours de parole

Comme évoqué mainte fois, la première étape du processus de la structuration en tours de parole d'un flux multilocuteur consiste à le découper en régions (segments), dont chacune est censée contenir la parole d'un unique locuteur. Une des méthodes les plus simples de segmentation est de découper d'une manière uniforme les intervalles de la parole trouvés par un détecteur de silence en segments de courtes durées (typiquement d'environ une seconde). Cette simple méthode de segmentation est traditionnellement connue dans le cadre de la structuration en tours de parole téléphonique, qui ont tendance à être très courts (Kenny, *et al.*, 2010b)(Shum, *et al.*, 2013). Une seconde étape de re-segmentation par l'algorithme de Viterbi est souvent appliquée comme traitement subséquent afin de raffiner les frontières entre les tours de parole.

8.2.2 I-vecteurs pour la représentation des tours de parole

Dans le contexte de la structuration en tours de parole, le segment vocal représentant un tour de parole est considéré comme l'unité fondamentale du problème ou ce qui est largement connu, dans le jargon de la reconnaissance des formes, par le terme *échantillon*. En outre, une agrégation des segments homogènes (c.-à-d. appartenant au même locuteur) dans un ensemble constitue une *classe* (appelée un *locuteur* dans notre cas). Ainsi, la structuration en tours de parole se réduit à un problème de regroupement automatique des tours de parole.

Dans le cadre de cette thèse, chaque tour de parole sera représenté par un vecteur dans l'espace des i-vecteurs de faible dimension (typiquement de dimension 100). En effet, la représentation de la parole par des i-vecteurs a prouvé son efficacité non seulement en reconnaissance du locuteur (Dehak, *et al.*, 2010)(Dehak, *et al.*, 2011a)(Dehak, *et al.*, 2011b)(Senoussaoui, *et al.*, 2013a) et en reconnaissance de la langue (Dehak, *et al.*, 2011c), mais aussi dans la structuration en tours de parole et le regroupement en locuteurs (Shum, *et al.*, 2011)(Shum, *et al.*, 2012)(Shum, *et al.*, 2013)(Senoussaoui, *et al.*, 2013b).

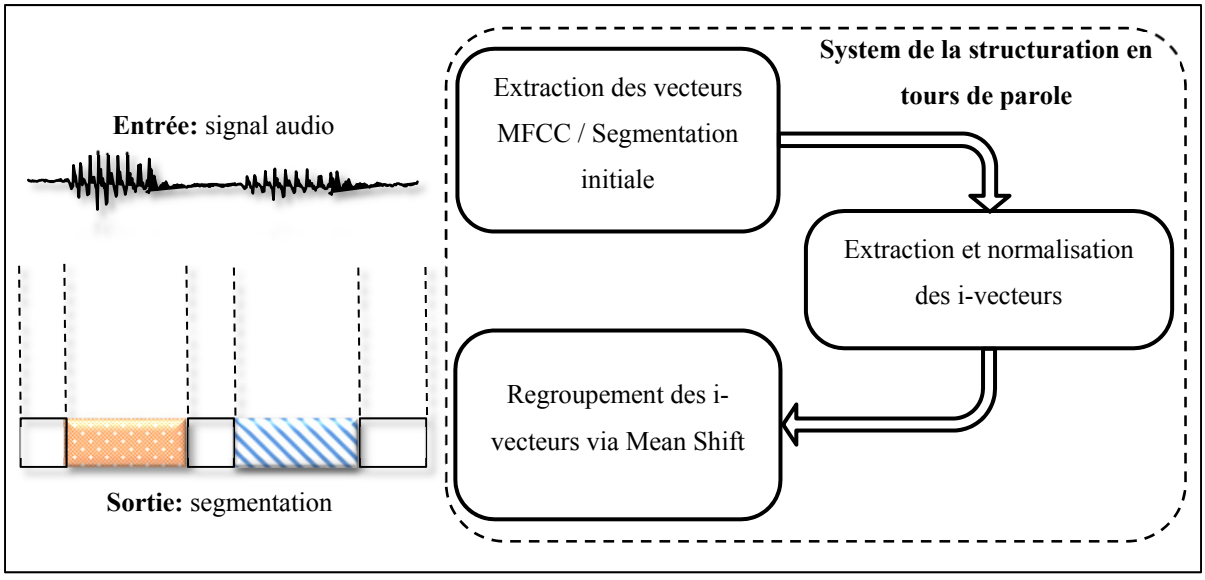

Figure 8.1 Illustration graphique de la structure générale d'un système de segmentation en tours de parole à base de la représentation en i-vecteurs et de regroupement via l'algorithme de Mean Shift.

8.2.3 Normalisation des i-vecteurs

De par sa nature, un i-vecteur modélise une vaste gamme de variabilités intrinsèques du signal vocal. Ainsi, il est toujours nécessaire de normaliser ces i-vecteurs selon des modalités qui varient d'une application à une autre. En nous basant sur la définition des notions de la *classe* et de ses *échantillons* relatives à notre problème (voir Section 8.2.2), nous présenterons dans les sous-sections suivantes quelques méthodes destinées à normaliser les i-vecteurs dans notre contexte de la structuration en tours de parole téléphonique.

8.2.3.1 Analyse en composantes principales (PCA)

Récemment, les auteurs de l'article (Shum, *et al.*, 2011) ont montré que la projection des i-vecteurs dans le sous-espace de PCA, formé par les axes ayant de fortes variances, compense la variabilité intrasession. De plus, les axes du sous-espace de PCA sont pondérés par la racine carrée des valeurs propres correspondant afin d'accentuer leur importance.

À vrai dire, ce traitement via la PCA est considéré comme local, et ce, du fait que la matrice de covariance de PCA est estimée localement pour chaque conversation en utilisant seulement ses propres segments (i-vecteurs). De par ce caractère local, la normalisation par la projection de PCA a l'avantage de ne solliciter aucun ensemble externe de données de développement.

Les auteurs de (Shum, *et al.*, 2011) recommandent de choisir la dimension de la PCA de manière à conserver une quantité de 50 % de la variance totale des données. Dans nos travaux, nous notons cette quantité par le caractère η.

Finalement, la normalisation[14] à **1** de la norme euclidienne des i-vecteurs est également souhaitable dans ce contexte tout comme dans le contexte de la vérification du locuteur (Garcia-Romero, 2011).

---

[14] http://people.csail.mit.edu/sshum/talks/tutorial_ISSPA_vPDF.pdf

8.2.3.2 Normalisation via l'inverse de la matrice de covariance intraclasse (WCCN)

La normalisation des i-vecteurs en utilisant l'inverse d'une matrice de covariance intraclasse est devenue une pratique courante dans le domaine de reconnaissance du locuteur (Dehak, *et al.*, 2010)(Dehak, *et al.*, 2011a)(Dehak, *et al.*, 2011b)(Senoussaoui, *et al.*, 2013a). L'idée derrière cette normalisation est de pénaliser les axes à forte variance intraclasse en appliquant une rotation de chaque observation (c.-à-d. i-vecteur dans notre cas) à l'aide d'une décomposition de l'inverse de la matrice de covariance intraclasse (Hatch, *et al.*, 2006).

8.2.3.3 Normalisation via la matrice de covariance interclasse (BCCN)

Par analogie avec l'approche de WCCN, nous proposons une nouvelle méthode de normalisation basée cette fois-ci sur l'accentuation de la variance interclasse des axes de notre espace. Pour ce faire, nous appliquons une rotation de chaque observation (c.-à-d. i-vecteur dans notre cas) via la décomposition de la matrice **B** de covariance interclasse. Cette matrice est donnée par la formule suivante :

$$\mathbf{B} = \frac{1}{n}\sum_{r=1}^{R} n_r (\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})' \tag{8.1}$$

où la somme s'étend sur les R conversations (chaque conversation contient n_r segments dont chacun est représenté par un i-vecteur) d'un ensemble d'apprentissage, $\bar{\mathbf{x}}_r = \frac{1}{n_r}\sum_{s=1}^{n_r} \mathbf{x}_r(s)$ est le i-vecteur moyen des segments de la conversation r et $\bar{\mathbf{x}} = \frac{1}{n}\sum_{r=1}^{n} \bar{\mathbf{x}}_r$ est le i-vecteur moyen de l'ensemble de n i-vecteurs (segments) de toutes les conversations.

8.2.4 Regroupement via le Décalage de la moyenne

Dans cette section, nous proposons deux solutions afin de remédier au problème de l'unique bande passante et celui de la surestimation du nombre de locuteurs. Effectivement, ces deux

problèmes constituent souvent un bémol aux mécanismes de regroupement à base de l'algorithme de *Décalage de la moyenne*.

8.2.4.1 Bande passante dépendante de conversation

Il est bien connu dans la littérature (Comaniciu, *et al.*, 2002) que l'une des limites de l'algorithme de *Décalage de la moyenne* est le besoin de fixer empiriquement une largeur de bande (bande passante) *h* qui servira dans la phase d'exploitation à la structuration de différents flux audio. L'utilisation d'une largeur de bande fixe n'est pas en général un choix convenable, et ce, du fait que la structure locale et le nombre des observations sont susceptibles d'être variables d'une conversation à une autre. Nous nous sommes aperçus que l'utilisation d'une largeur de bande qui varie d'une conversation à une autre est avantageuse dans le cas de la structuration en tours de parole à base de l'algorithme MS. Afin de tenir compte de la disparité causée par la durée variable des conversations (c.-à-d. un nombre variable de segments d'une conversation à une autre), nous adoptons une forme de la largeur de bande variable proposée dans (Stafylakis, *et al.*, 2012). Cette forme est conçue principalement pour des fins de lissage de l'estimateur de la fonction de densité (voir équation 6.1) dans le cas des conversations de courtes durées.

La largeur de bande $\tilde{h}_r$ dépendante d'une conversation *r* est contrôlée par deux variables, à savoir, un facteur d'ajustement τ et la bande passante fixe *h* (c.-à-d. la bande indépendante de conversation) :

$$\tilde{h}_r = 1 - \left(\frac{n_r \tau (1-h)}{n_r \tau + (1-h)} \right) \tag{8.2}$$

où n_r est le nombre de segments de la conversation *r* à structurer. Notons que $\tilde{h}_r \geq h$ avec une égalité si et seulement si le nombre n_r des segments est très grand.

8.2.4.2 Élagage des classes éparses

Un des points faibles de l'algorithme de *Décalage de la moyenne* est l'absence de contrôle concernant le nombre d'observations affectées à une classe détectée, ce qui produit parfois des classes avec un très petit nombre d'observations. Afin de contourner cette tendance, nous proposons une simple stratégie d'élagage. Cette stratégie consiste à désigner les groupes contenant un faible nombre d'échantillons (inférieure ou égale à une constante p), ces groupes seront par la suite fusionnés avec leurs groupes voisins les plus proches.

8.3 Expérimentation

Une étude expérimentale approfondie sera présentée dans cette section montrant l'efficacité de notre nouvelle version d'algorithme de *Décalage de la moyenne* une fois testée sur la tâche de la structuration en tours de parole téléphonique.

8.3.1 Détails d'implémentation

Dans cette section, nous présentons les détails nécessaires permettant la reproduction authentique des résultats obtenus.

8.3.1.1 Corpus CallHome des données téléphoniques

Tel qu'introduit brièvement dans la Section 1.4.2.1, le corpus de données CallHome (Martin, et al., 2001) se divise en deux sous-ensembles disjoints (voir Figure 8.2), l'un pour le développement et l'autre pour le test. Nous utilisons le premier sous-ensemble à des fins d'estimation d'hyper-paramètres (c.-à-d. la bande passante h, la quantité η de variabilité retenue par la PCA et le facteur d'élagage p) ainsi que pour le choix de meilleure configuration de la normalisation des i-vecteurs (c.-à-d. la meilleure combinaison des méthodes de normalisation). Le sous-ensemble du test sera utilisé pour valider la capacité de généralisation de nos algorithmes. Ce sous-ensemble sera utilisé également pour optimiser à

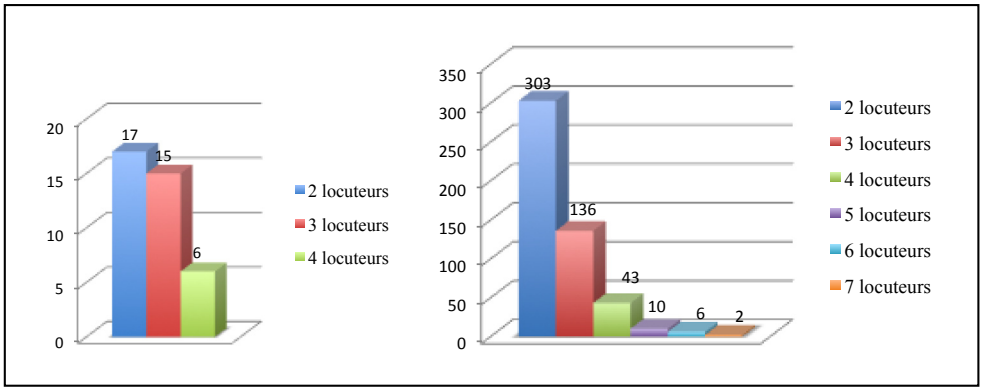

Figure 8.2 Les deux sous-ensembles du corpus de CallHome (à gauche l'ensemble du développement et à droite l'ensemble du test) tels qu'ils sont représentés en fonction des groupes représentant le nombre des locuteurs impliqués dans un enregistrement. Les chiffres en ordonnée représentent le nombre d'enregistrements.

nouveau l'hyper-paramètre *h* afin de pouvoir comparer nos résultats avec ceux de l'état de l'art (Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Vaquero, 2011)(Shum, *et al.*, 2012).

La largeur de bande *h* représente l'hyper-paramètre scalaire le plus important qui influence les performances et qui sollicite une estimation à partir d'un ensemble de développement indépendant. Ainsi, l'algorithme de MS ne représente pas un grand risque de surapprentissage.

Finalement, en plus du fait que les deux sous-ensembles de CallHome sont complètement disjoints, il est important de souligner qu'il existe d'autres sources de disparité entre ces deux sous-ensembles, compliquant davantage la tâche d'optimisation des hyper-paramètres (voir Figure 8.2). En guise d'exemple, les conversations sont en différentes langues (six langues en tout), les durées des enregistrements sont variables (entre une à dix minutes), le nombre maximum des locuteurs impliqués dans les conversations de développement (quatre locuteurs) est inférieur à celui des conversations du test (sept locuteurs) et enfin, le nombre total des conversations de développement est relativement faible (38 conversations).

8.3.1.2 Extraction des i-vecteurs

- ***Paramétrage du signal vocal*** : toutes les 10*ms*, 20 coefficients cepstraux de *Mel* (MFCC) sont extraits à partir d'une fenêtre d'analyse de *Hamming* de taille de 25*ms* (19 coefficients de MFFC + un coefficient d'énergie). Dans le contexte de la structuration en tours de parole, aucune méthode de normalisation à court terme des paramètres n'est appliquée (Reynolds, *et al,*. 2000) tel qu'expliquée dans la Section 2.1.1.
- ***Modèle du monde (UBM)*** : nous utilisons un modèle UBM indépendant du genre contenant 512 gaussiennes. Cet UBM est entrainé à partir des données d'apprentissage de NIST suivantes : les données Switchboard II de LDC, phases 2 et 3 ; Switchboard de cellulaire, parties 1 et 2, et les données « MIX » de NIST SRE 2004 à 2005 (uniquement de la parole téléphonique).
- ***Extracteur des i-vecteurs*** : nous utilisons un extracteur des i-vecteurs indépendant du genre de dimension 100. Les paramètres de cet extracteur des i-vecteurs sont estimés à partir des mêmes données que celles utilisées pour l'entrainement d'UBM en plus des données du corpus de Fisher.
- ***Normalisation des i-vecteurs*** : parmi les méthodes de normalisation présentées dans la Section 8.2 (la PCA et la normalisation à **1** de la norme euclidienne, la WCCN et la BCCN), seuls les paramètres de WCCN et de BCCN nécessitent un ensemble d'entrainement indépendant. Afin de les estimer, nous utilisons la parole téléphonique des enregistrements « MIX » (chaque enregistrement contient environ 2,5 minutes de la parole d'un unique locuteur) de NIST SRE 2004 et 2005. En nous basant sur la segmentation fournie par un détecteur de la parole pour un enregistrement donné, nous extrayons un i-vecteur de dimension 100 pour chaque segment.

8.3.1.3 Protocole d'évaluation

Afin d'évaluer les performances des différents systèmes de structuration en tours de parole, nous adoptons le taux d'erreur de la structuration proposé comme mesure principale par NIST (*Diarization Error Rate*, DER) (voir Section 1.4.2). Pour un ensemble de conversations à

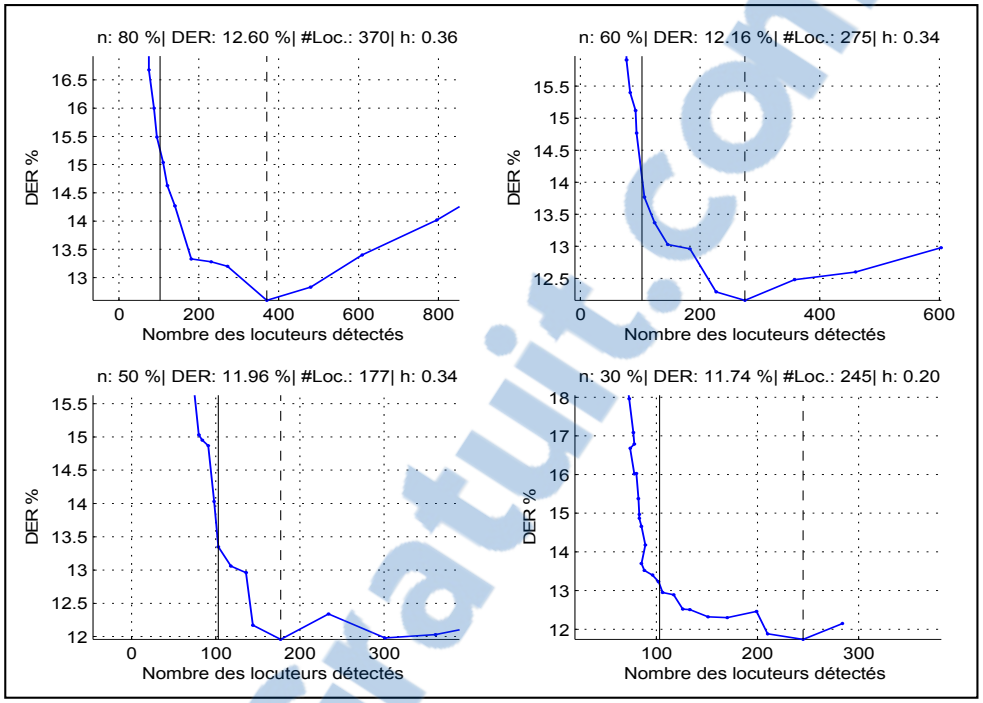

Figure 8.3 Résultats (DER/nombre des locuteurs estimés) de la stratégie totale de regroupement (STR) obtenus à partir de l'ensemble de développement en utilisant la normalisation des i-vecteurs via la PCA. Le minimum de DER, sa bande passante fixe correspondante (h) ainsi que son nombre de locuteurs estimés correspondant (#Loc.) sont également fournis pour chaque facteur $\eta = 80$, 60, 50 et 30 de la PCA.

structurer, le DER est évalué en utilisant le script d'évaluation de NIST (md-eval-v21.pl[9]) et un seul fichier « .rttm » obtenu par la concaténation des fichiers « .rttm » associés à chaque conversation. Lors d'évaluation du DER, il est courant, dans le cas de la parole téléphonique, d'ignorer les chevauchements de la parole et de tolérer également une marge de 250*ms* pour les erreurs de localisation des frontières entre les tours de parole.

Tel que défini dans la Section 1.4.2, le nombre des locuteurs détectés (*Number of Detected Speakers,* NDS) fournit aussi une mesure efficace des performances des systèmes de la structuration en tours de parole. Le NDS et sa moyenne (Average of the *Number of Detected Speakers,* ANDS) calculée sur l'ensemble des conversations à structurer seront également considérés comme deux mesures d'évaluation supplémentaires dans cette étude. Finalement, nous adoptons l'illustration graphique du DER en fonction de la NDS (voir Section 1.4.2) pour mieux représenter les comportements des systèmes développés sur une vaste plage des

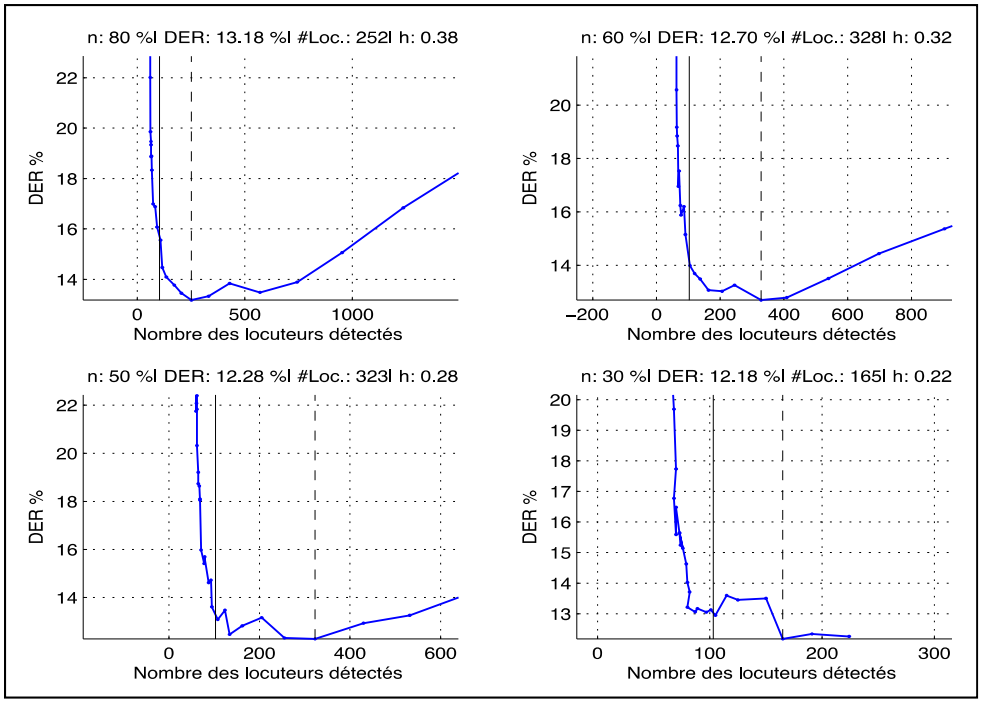

Figure 8.4 Résultats (DER/nombre des locuteurs estimés) de la stratégie sélective de regroupement (SSR) obtenus à partir de l'ensemble de développement en utilisant la normalisation des i-vecteurs via la PCA. Le minimum de DER, sa bande passante fixe correspondante (*h*) ainsi que son nombre de locuteurs estimés correspondant (#Loc.) sont également fournis pour chaque facteur η = 80, 60, 50 et 30 de la PCA.

configurations obtenu par la variation de la bande passante *h* de l'algorithme de *Décalage de la moyenne*.

8.3.2 Résultats et discussions

Dans cette section, nous présenterons et discuterons les résultats de notre étude expérimentale des effets des méthodes de normalisation des i-vecteurs (voir Section 8.2.2) sur les mécanismes de regroupement à base de l'algorithme de MS.

8.3.2.1 Optimisation des hyper-paramètres à partir des données de développement

Dans le but d'établir nos systèmes de référence, nous commençons la présente étude par la confirmation du choix de retenir 50 % de la variabilité totale des segments d'une

conversation donnée. Pour ce faire, nous avons exécuté maintes fois les deux stratégies de regroupement, *totale* (STC) et *sélective* (SSC), de l'algorithme de *Décalage de la moyenne* à base de la distance de cosinus (voir Figures 8.3 et 8.4), et ce, en faisant varier la bande passante (sur une importante plage de points entre 0 et 2) et en changeant le facteur η de la variabilité retenue par la PCA (η = 80, 60, 50 et 30).

En observant les Figures 8.3 et 8.4, nous constatons que les résultats pour les deux stratégies obtenus avec le facteur η = 30 % sont légèrement meilleurs que ceux obtenus avec η = 50 %. Or, les graphes sont irréguliers dans le premier cas (c.-à-d. η = 30 %), ce qui nous incite à favoriser η = 50 % tout comme dans l'article (Shum, *et al.*, 2011). De plus, il faut noter que pour toutes les configurations, le minimum du taux d'erreur DER est atteint avec une surestimation du nombre de locuteurs (la ligne verticale continue représente le vrai nombre de locuteurs et la ligne discontinue représente celui atteint avec le minimum de DER). Heureusement, la surestimation est souvent préférable à la sous-estimation, car la deuxième peut être remédiée par un élagage de classes éparses (voir Section 8.2.4.2).

– ***Impact de la normalisation à 1 de la norme euclidienne*** :

Après avoir établi le système de référence, nous commençons par la vérification de l'effet de la normalisation à **1** de la norme euclidienne des i-vecteurs avant l'estimation et la projection dans l'espace réduit de la PCA. D'une façon étonnante, cette simple opération a amélioré la DER de 2 % en valeur absolue (voir la troisième ligne -Norme. Euc.- du Tableau 8.1), la DER a diminué de 11,9 % (voir Figure 8.3) à 10 % pour la *stratégie totale* et de 12,2 % (voir. Figure 8.4) à 10,2 % pour la *stratégie sélective*. En outre, dans le cas de la *stratégie sélective*, le nombre des locuteurs détectés (NDS) diminue de 323 à 281, se rapprochant ainsi de la valeur réelle qui est de 103 locuteurs. Toutefois, dans le cas de la *stratégie totale* le NDS augmente de 177 à 316 locuteurs. Dès lors, ce système sera considéré comme notre nouveau système de référence.

– ***Impact de la normalisation WCCN***

Dans cette expérience, nous normalisons d'abord les i-vecteurs de chaque conversation en utilisant la décomposition de l'inverse de la matrice de covariance intraclasse (voir Section 8.2.3.2). Ensuite, ces i-vecteurs subissent les mêmes transformations que celles de

l'expérience précédente, à savoir, la normalisation à **1** des normes euclidiennes suivie par la projection dans l'espace de la PCA représentant 50 % de la variabilité totale.

En observant la quatrième ligne du Tableau 8.1, il s'avère que la WCCN entraîne une dégradation des performances par rapport à l'expérience précédente (le DER est passé de 10 % à 11,7 % dans le cas de la *stratégie totale* et de 10,2 % à 11,7 % dans le cas de la *stratégie sélective*). Or, ces résultats ne s'accordent pas avec nos espérances du fait que la WCCN était très efficace dans le cas de la reconnaissance du locuteur. Ce comportement est probablement dû à une interaction entre la PCA et la WCCN.

– ***Impact de la normalisation BCCN*** :

Mis à part l'utilisation de la matrice de covariance interclasse au lieu de l'inverse de la matrice de covariance intraclasse, la procédure de la BCCN est exactement la même que celle de la WCCN.

Tableau 8.1 Résultats (DER, NDS) obtenus à partir des données de développement illustrant l'effet de diverses méthodes de normalisation. *h* est la bande passante correspondante au minimum de DER et *p* est le facteur d'élagage. Le nombre réel de locuteurs est de 103.

	Stratégie *totale* de MS				Stratégie *sélective* de MS			
	DER (%)	NDS	*h*	*p*	DER (%)	NDS	*h*	*p*
Norme. Euc.	10.0	316	0.34	0	10.2	281	0.34	0
WCCN	11.7	320	0.30		11.7	343	0.28	
BCCN	7.6	285	0.26		7.7	189	0.28	
Var. *h*	7.5	300	0.22		7.6	203	0.24	
Élagage	8.3	109	**0.32**	1	7.5	111	0.24	3

Dans la cinquième ligne (BCCN) du Tableau 8.1, nous constatons deux améliorations remarquables par rapport aux résultats de notre système de référence (voir la troisième ligne -Norme. Euc- du Tableau 8.1). D'une part, nous avons obtenu une bonne réduction de DER qui passe de 10 % à 7,6 % dans le cas de la *stratégie totale* et de 10,2 % à 7,7 % dans le cas de la *stratégie sélective*. D'autre part, nous constatons également que le nombre des locuteurs détectés (NDS) est devenu plus proche de la valeur réelle (103 locuteurs), et ce, notamment

pour la *stratégie sélective* (NDS = 189 locuteurs). Des illustrations graphiques montrant les effets positifs de l'application successive des méthodes de la normalisation des i-vecteur dans le contexte de l'algorithme de MS sont fournies à l'Annexe II.

– ***Bande passante dépendante de conversation*** :

Dans cette expérience, nous avons repris la même configuration que l'expérience précédente (c.-à-d. la BCCN suivie par la normalisation à **1** de la norme euclidienne et par la projection de la PCA) à l'exception de l'adoption de la forme variable de la bande passante (voir équation 8.2).

Une amélioration négligeable du DER (voir la sixième ligne -Var. *h*- du Tableau 8.1) est notée par rapport aux résultats de l'expérience précédente.

– ***Élagage des classes éparses*** :

Bien que nous ayons réussi à réduire le DER d'environ 12 % à environ 7 % pour les deux stratégies du regroupement, l'estimation du nombre de locuteurs (NDS) correspondant au minimum de DER reste encore supérieure à la valeur réelle (103 locuteurs). Tel qu'expliqué dans la section 8.2.4.2, notre stratégie d'élagage consiste à sélectionner les groupes épars (c.-à-d. les classes dont le nombre des échantillons est inférieur ou égal à une constante p) afin de les fusionner avec leurs groupes voisins les plus proches.

Les résultats correspondants apparaissent dans la dernière ligne du Tableau 8.1 (Élagage). Nous observons que pour la *stratégie totale*, la fusion des groupes ayant une seule observation (p = 1) réduit le NDS de 300 à 109, tandis que le DER augmente légèrement de 7,5 % à 8,3 %. Pour la *stratégie sélective*, nous avons obtenu une amélioration remarquable du NDS (111 locuteurs au lieu de 203) avec un facteur p = 3, le DER demeure quasiment inchangé passant de 7,6 % à 7,5 %.

8.3.2.2 Résultats obtenus à partir de l'ensemble du test

Nous poursuivons cette étude en réalisant une série d'expériences qui vise à explorer la capacité de la généralisation de nos algorithmes. Pour ce faire, nous avons exécuté les deux

stratégies de MS en utilisant les données du test (voir Figure 8.2) et les paramètres (c.-à-d. la bande passante *h* ainsi que le facteur *p* en cas d'élagage) déjà estimés à partir de l'ensemble de développement. Le Tableau 8.2 regroupe les résultats les plus importants. Le terme « Fix. *h* » à la ligne 3 du tableau se réfère au meilleur système utilisant une bande passante fixe (c.-à-d. le système dont les résultats sont présentés à la ligne 5 (BCCN) du Tableau 8.2). Dans ce système, on a utilisé respectivement la BCCN suivie par la normalisation à **1** de la norme euclidienne et une projection par la PCA avec η = 50 %. Les résultats présentés dans la quatrième ligne du Tableau 8.2 (Var. *h*) sont produits par un système exactement semblable au précédent (c.-à-d. le système Fix. *h*) à l'exception de l'utilisation d'une largeur de bande variable (voir équation 8.2). Enfin, la dernière ligne du Tableau 8.2 (Élagage) met en évidence l'utilité de l'opération d'élagage appliquée sur le système à largeur de bande variable (Var. *h*).

Tableau 8.2 Résultats obtenus à partir des données du test en utilisant les paramètres (*h*, *p*) optimisés à partir de l'ensemble de développement. Le nombre réel des locuteurs dans le corpus du test est de 1283.

	Stratégie *totale* de MS				Stratégie *sélective* de MS			
	DER (%)	NDS	*h*	*p*	DER (%)	NDS	*h*	*p*
Fix. *h*	14.3	3456	0.26	0	13.9	3089	0.28	0
Var. *h*	12.7	2550	0.22	0	12.6	2310	0.24	0
Élagage	**12.4**	**1361**	**0.32**	1	14.3	1501	0.24	3

L'efficacité de la largeur de bande variable dans la réduction du DER est bien perçue à travers les résultats (les DER passent de 14,3 % à 12,7 % dans le cas de la *stratégie totale* et de 13,9 % à 12,6 % dans le cas de la *stratégie sélective*). Nous observons également que le NDS est réduit de 3456 à 2550 dans la *stratégie totale* et de 3089 à 2310 dans le cas de la *stratégie sélective* (voir les lignes 3 - Fix. *h* - et 4 - Var. *h* - du Tableau 8.2). À l'inverse de ce qui est observé sur l'ensemble de développement, l'opération d'élagage mène à une dégradation du DER de 12,6 % à 14,3 % dans le cas de la *stratégie sélective*. Or, dans le cas de la *stratégie totale*, cette même opération a été énormément efficace au point d'obtenir le minimum de DER (12,4 %) de l'ensemble du test et celui de l'ensemble du développement

(8,3 %) avec la même largeur de bande (voir les dernières lignes du Tableau 8.1 et du Tableau 8.2). Finalement, ces résultats discutés ci-dessus confirment la capacité de généralisation des deux stratégies de regroupement de *Décalage de la moyenne* à base de la distance de cosinus.

Parmi toutes les publications rapportant des résultats sur l'ensemble des données CallHome (Martin, et al., 2001)(Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Vaquero, 2011)(Shum, *et al.*, 2012), seule la thèse de Vaquero présente les résultats par un DER total calculé à partir de l'ensemble de tous les fichiers (Vaquero, 2011). Dans sa thèse, Vaquero a utilisé des facteurs de locuteurs (*speaker factors*) (Kenny *et al.,* 2008), plutôt que des i-vecteurs, pour représenter les segments de parole. Pour le regroupement, l'auteur a adopté un système à plusieurs couches basé principalement sur la classification hiérarchique (HAC), l'algorithme des k-moyennes et la segmentation de Viterbi. Il a également estimé ces hyper-paramètres à partir d'un ensemble de développement indépendant composé uniquement des enregistrements à deux locuteurs.

Afin d'obtenir un DER de 13,7 % sur l'ensemble du test de CallHome, il a été contraint de fournir le nombre réel des locuteurs comme un critère d'arrêt à son algorithme de regroupement hiérarchique, autrement, son DER a été de 19,8 %. Ainsi, nous sommes en mesure d'obtenir une amélioration relative de 37 % de DER total (12.4 %) par rapport à ces résultats.

8.3.2.3 Résultats regroupés en fonction du nombre de locuteurs

Dans cette section, nous procédons à la comparaison de nos résultats avec celles de l'état de l'art rapportés dans les articles suivants (Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Shum, *et al.*, 2012). Pour ce faire, nous devons adopter la même convention de la présentation des résultats de la structuration en tours de paroles. Les résultats de ces travaux sont présentés avec six catégories, dont chacune regroupe toutes les conversations impliquant le même nombre de locuteurs. Ainsi, un taux d'erreur (DER) et une moyenne du nombre des locuteurs détectés (ANDS) sont calculés pour chacune des six catégories (voir Figure 8.5). De plus, les

résultats de ces travaux sont obtenus en optimisant les hyper-paramètres à partir des mêmes données du test. Ceci est probablement dû au fait que l'ensemble de développement contient seulement des conversations ayant entre deux et quatre locuteurs.

Pour chaque stratégie de regroupement (c.-à-d. *totale* et *sélective*) et chaque forme de bande passante (c.-à-d. fixe et variable), nous avons réalisé deux types d'expériences. Quant au premier type d'expérience, nous avons utilisé les hyper-paramètres optimisés sur l'ensemble des données de développement et quant au deuxième, nous avons utilisé les hyper-paramètres optimisés sur l'ensemble des données du test. Les résultats des stratégies *totale* et *sélective* sont respectivement présentés par les Tableaux 8.3 et 8.4. En effet, l'aspect de la généralisation de nos deux stratégies de regroupement est clairement perçu à partir de ces résultats, et ce, notamment dans le cas de la *stratégie totale* adoptant la forme fixe de la bande passante (voir Tableau 8.3).

Tableau 8.3 Résultats de la stratégie totale de regroupement obtenus à partir des données du test en utilisant deux types d'hyper-paramètres (optimisés respectivement à partir des données de développement/du test). Les résultats sont représentés par catégories de nombre de locuteurs. Pour des raisons de simplicité, la colonne grise fournie les numéros des lignes.

		Nombre de locuteurs	2	3	4	5	6	7	*h/p*
Paramètres dev.	2	DER (%)	11.9	13.5	15.6	22.9	29.5	28.4	**Fix. *h* 0.26 / 0**
	3	ANDS	6.5	7.2	7.9	8.5	9.6	11.0	
	4	DER (%)	7.8	12.5	16.2	23.1	30.4	28.6	Var. *h* 0.22 / 0
	5	ANDS	4.0	5.5	8.4	9.3	14.3	20.0	
Paramètres test	6	DER (%)	11.9	13.5	15.6	22.9	29.5	28.4	**Fix. *h* 0.26 / 0**
	7	ANDS	6.5	7.2	7.9	8.5	9.6	11.0	
	8	DER (%)	8.1	12.5	15.5	23.2	27.5	29.0	Var. *h* 0.20 / 0
	9	ANDS	4.2	6.1	10.5	11.9	14.6	24.0	

Par l'observation des résultats présentés au Tableau 8.3, nous constatons tout d'abord que *la stratégie totale* de regroupement ne sollicite aucune opération d'élagage (voir $p = 0$ dans toutes les cellules de la dernière colonne du Tableau 8.3). De plus, en ce qui concerne le DER, les résultats obtenus via une bande passante fixe (Fix. *h*) sont comparables à ceux

obtenus via la bande passante variable (Var. *h*). Or, les estimations du nombre de locuteurs dans le premier cas (c.-à-d. Fix. *h*) sont plus proches des chiffres réels que dans le cas de la bande passante variable (voir les lignes 3 et 7 par rapport aux lignes 5 et 9 du Tableau 8.3). D'une manière générale, la largeur de la bande variable contribue à réduire le DER et le ANDS pour les conversations impliquant un faible nombre de locuteurs (2 ou 3 locuteurs).

À l'inverse de la *stratégie totale*, les résultats représentés dans le Tableau 8.4 montrent que l'élagage des classes éparses est une opération nécessaire dans le cas de la *stratégie sélective*. La combinaison d'une largeur de la bande variable avec l'élagage des groupes ayant au maximum trois observations (c.-à-d. $p = 3$) a permis l'obtention de meilleurs résultats, tant pour les DER que pour les ANDS (voir les lignes 4 et 5 et les lignes 8 et 9 dans le Tableau 8.4). En ce qui concerne les taux d'erreurs DER, les résultats de la *stratégie totale* (voir Tableau 8.3) sont semblables à ceux de la *stratégie sélective* (voir Tableau 8.4) ; or les résultats de cette dernière sont meilleurs en tenant compte du nombre moyen des locuteurs détectés (ANDS).

Tableau 8.4 Résultats de la stratégie sélective de regroupement obtenus à partir des données du test en utilisant deux types d'hyper-paramètres (optimisés respectivement à partir des données de développement/du test). Les résultats sont représentés par catégories de nombre de locuteurs. Pour des raisons de simplicité, la colonne grise fournie les numéros des lignes.

		Nombre de locuteurs	2	3	4	5	6	7	*h/p*
Paramètres dev.	2	DER (%)	9.9	12.6	15.5	22.6	29.3	29.9	Fix. 0.28 / 3
	3	ANDS	2.8	3.2	3.3	3.8	4.0	6.0	
	4	DER (%)	7.2	13.1	15.6	22.8	29.0	27.7	Var. 0.24 / 3
	5	ANDS	2.3	2.9	3.4	3.5	4.1	5.5	
Paramètres test	6	DER (%)	10.8	12.8	15.6	21.7	26.7	27.3	Fix. 0.26 / 3
	7	ANDS	2.9	3.3	3.7	3.9	4.3	6.5	
	8	DER (%)	8.1	12.6	15.9	22.2	26.1	27.6	Var. 0.18 / 3
	9	ANDS	2.4	3.3	4.8	5.2	5.8	10.0	

8.3.2.4 Resegmentation de Viterbi

L'ajustement des frontières entre les tours de parole de différents locuteurs en utilisant l'algorithme de recherche de chemin optimal de Viterbi est une procédure standard adoptée afin d'améliorer les performances des systèmes de la structuration en tours de paroles (Gupta, *et al.*, 2007)(Kenny, *et al.*, 2010b). L'idée consiste à utiliser les vecteurs acoustiques afin de construire un modèle (GMM, par exemple) pour chaque locuteur détecté par la première passe du système de la structuration en tours de parole. Ces modèles sont ensuite considérés comme les états d'un modèle de Markov caché (*Hidden Markov Model*, HMM). L'algorithme de Viterbi est finalement utilisé pour la réaffectation des vecteurs acoustiques au locuteur le plus probable.

Les résultats présentés dans le Tableau 8.5 montrent clairement l'efficacité de cette pratique dans la réduction des taux d'erreurs DER de la structuration à base de nos deux stratégies de regroupement de *Décalage de la moyenne*. Il est à souligner que les résultats obtenus sans la resegmentation de Viterbi (voir les cellules grises dans le Tableau 8.5) sont les mêmes résultats présentés dans les sixièmes et huitièmes lignes des Tableaux 8.3 et 8.4.

Tableau 8.5 Résultats obtenus à partir des données de l'ensemble du test après une resegmentation de Viterbi (les hyper-paramètres sont estimés à partir des mêmes données du test).

		Nombre de locuteurs	2	3	4	5	6	7
Totale MS	Fix. *h*	- Viterbi	11.9	13.5	15.6	22.9	29.5	28.4
		+ Viterbi	11.2	12.3	14.5	22.8	27.3	27.4
	Var. *h*	- Viterbi	8.1	12.5	15.5	23.2	27.5	29.0
		+ Viterbi	7.6	11.8	14.6	23.5	26.5	28.4
Sélective MS	Fix. *h*	- Viterbi	10.8	12.8	15.6	21.7	26.7	27.3
		+ Viterbi	10.1	11.6	14.3	22.0	25.9	26.7
	Var. *h*	- Viterbi	8.1	12.6	15.9	22.2	26.1	27.6
		+ Viterbi	7.5	11.8	14.9	22.8	25.9	26.9

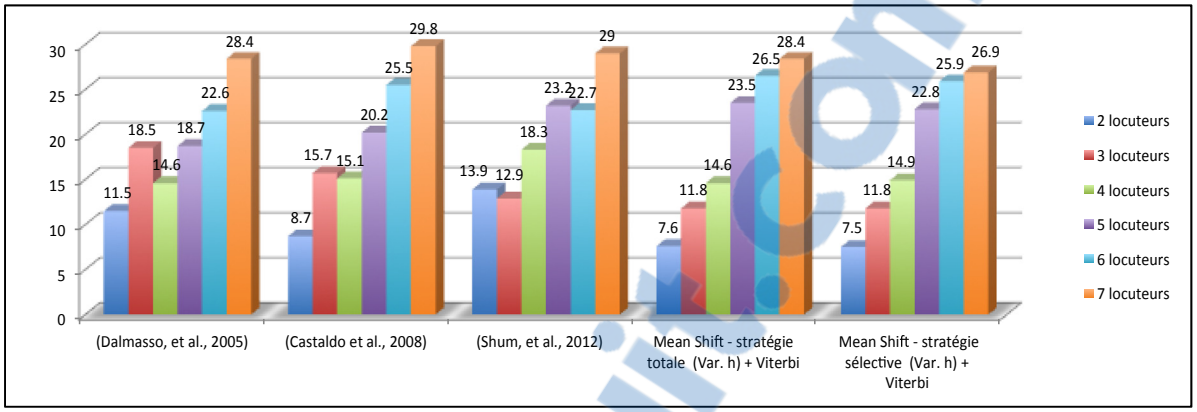

Figure 8.5 Comparaison des résultats (DER % en ordonnée) des deux versions, totale (STR) et sélective (SSR), de l'algorithme de Mean Shift avec les résultats de l'état de l'art obtenus à partir des mêmes données du test du corpus CallHome.

8.3.2.5 Comparaison avec les résultats de l'état de l'art

À ce stade, nous sommes en mesure de mener une comparaison entre nos résultats et ceux de l'état de l'art (Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Shum, *et al.*, 2012). Or, suite à plusieurs facteurs, une comparaison parfaite reste difficile. Contrairement à (Vaquero, 2011) et à notre travail, les auteurs de (Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Shum, *et al.*, 2012) n'ont pas utilisé un ensemble indépendant des données de développement afin d'estimer les hyper-paramètres. En outre, dans les travaux de (Dalmasso, *et al.*, 2005) et (Castaldo *et al.*, 2008), les auteurs adoptent des hypothèses sur le nombre maximum de locuteurs pouvant être présents dans un segment de parole. Finalement, dans le travail de (Shum, *et al.*, 2012) les auteurs estiment le nombre de locuteurs dans une conversation en une étape séparée de celle du regroupement.

Afin de faciliter la tâche de comparaison, nous avons représenté graphiquement dans la Figure 8.5 les résultats (en pourcentage de DER) de nos meilleures configurations (c.-à-d. les résultats présentés dans les cinquièmes et neuvièmes lignes du Tableau 8.5) et ceux de (Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Shum, *et al.*, 2012). En observant ces graphiques, il s'avère que nos résultats sont comparables à ceux de l'état de l'art.

Tableau 8.6 Comparaison avec les résultats de Dalmasso selon le critère du nombre moyen des locuteurs détectés (ANDS).

Nombres réels de locuteurs	2	3	4	5	6	7
(Dalmasso *et al.* 2005)	1.9	2.3	3.3	4.4	4.8	6.5
Sélective MS	2.4	3.3	4.8	5.2	5.8	10.0

En outre, une comparaison à base du nombre moyen des locuteurs détectés (ANDS) n'est possible qu'avec le travail de (Dalmasso, *et al.*, 2005). Le Tableau 8.6 regroupe nos meilleurs résultats en terme de l'ANDS (voir la dernière ligne du Tableau 8.4) et ceux de (Dalmasso, *et al.*, 2005). D'une manière générale, les résultats sont comparables. Or, l'algorithme de *Décalage de la moyenne* a tendance à surestimer le nombre des locuteurs, ce qui est souvent préférable à la sous-estimation de ces derniers.

8.3.2.6 Temps d'exécution des algorithmes

L'évaluation du temps de calcul de nos algorithmes n'est guère une préoccupation dans cette étude. Toutefois, nous avons préféré fournir un minimum d'informations à ceux qui s'y intéressent. La Figure 8.6 illustre graphiquement la différence entre les deux stratégies de

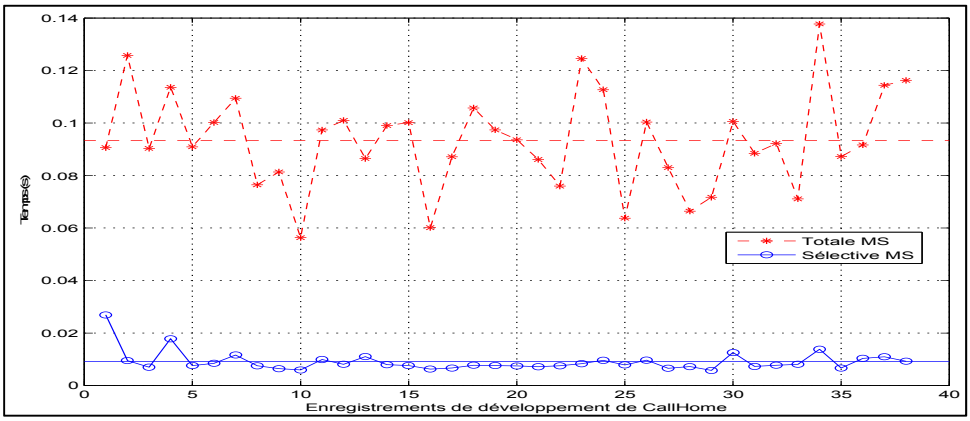

Figure 8.6 Comparaison du temps d'exécution (en secondes) des deux stratégies, totale (STR) et sélective (SSR), de l'algorithme de Mean Shift.

regroupement *totale* et *sélective* à cet égard. Le temps moyen pour segmenter une conversation est de 0,0934 seconde dans le cas de *la stratégie totale* contre seulement 0,0091 seconde dans le cas de la *stratégie sélective*.

CONCLUSION

Depuis toujours, reconnaître l'identité des personnes constitue un besoin fondamental. Toutefois, lorsqu'on ne dispose que de la voix comme caractéristique physique d'une personne, la reconnaissance automatique de son identité ne devient possible qu'à travers les systèmes de la reconnaissance du locuteur. Dans le contexte actuel, le développement des technologies à base de modules intelligents, capables d'identifier les gens à partir de leurs propres voix, est en plein essor.

Adoptées par l'institut américain NIST, les recherches en reconnaissance du locuteur ont connu d'énormes progrès au cours des deux dernières décennies. En effet, l'espace de la représentation de la voix par des vecteurs de faibles dimensions, dénommés i-vecteurs (Dehak, *et al.*, 2009)(Dehak, *et al.*, 2011b), est l'un de ces remarquables progrès, voire même le plus influent à nos jours. En outre, les classificateurs de nature probabiliste, tels que l'analyse discriminante linéaire probabiliste (PLDA) (Kenny, 2010a), sont devenus prédominants dans l'application des méthodes des i-vecteurs à la reconnaissance du locuteur. Toutefois, les classificateurs simples à base de la similarité angulaire du cosinus (SAC) restent compétitifs (Dehak, *et al.*, 2010)(Senoussaoui, *et al.*, 2013a). Dans l'ère actuelle des i-vecteurs, les recherches en reconnaissance du locuteur (notamment en vérification), telles qu'elles sont encadrées par NIST, ont atteint un degré de maturité très avancé.

Le principal objectif de cette thèse est d'améliorer la robustesse des systèmes de reconnaissance du locuteur opérant dans l'espace des i-vecteurs. Certes, la robustesse de ces systèmes est menacée par plusieurs facteurs une fois exploités dans les environnements réels. La première partie de cette thèse est consacrée au traitement de deux facteurs fondamentaux dégradant les performances des systèmes de vérification du locuteur, à savoir, le changement radical du canal de transmission (téléphone/microphone) ainsi que la différence physiologique entre les voix des hommes et celles des femmes. Dans la deuxième partie, nous avons élargi notre intérêt de recherche en entamant la question du regroupement automatique en locuteurs des segments vocaux.

Afin de remédier au problème du changement radical des canaux, nous avons proposé en premier temps une nouvelle architecture d'extraction des i-vecteurs convenables à la parole téléphonique comme à la parole microphonique. Cette architecture consiste à entrainer séparément deux matrices de la variabilité totale, l'une à partir des données téléphoniques et l'autre à partir des données microphoniques (Senoussaoui *et al.*, 2010). Ces deux matrices sont concaténées pour enfin produire la matrice de la variabilité totale d'extracteur des i-vecteurs indépendants du canal. Par la suite, cette architecture a été améliorée par l'introduction d'une étape du traitement subséquente réalisant une réduction supervisée de la dimensionnalité des i-vecteurs indépendants du canal (Dehak, *et al.*, 2011a)(Sturim, *et al.*, 2011)(Senoussaoui *et al.*, 2011b). La découverte de l'avantage de la réduction supervisée de la dimensionnalité nous a incités à revenir à l'exploration d'une simple idée antérieurement écartée. L'idée en question consiste à entrainer un extracteur des i-vecteurs à partir d'un ensemble regroupant les données téléphoniques et microphoniques (Senoussaoui *et al.*, 2011a). Ces deux architectures d'extracteur des i-vecteurs ont permis l'obtention des résultats de l'état de l'art des systèmes de la vérification du locuteur évalués selon les critères de NIST (environ 2 % d'EER pour la parole téléphonique et 3 % pour la parole microphonique). Et ceci, sans aucune sollicitation de l'information sur le type du canal utilisé à l'enregistrement/transmission de la voix des locuteurs.

Afin de pallier le problème de la dépendance du genre des systèmes de vérification du locuteur, nous avons proposé deux solutions au niveau des classificateurs afin de rendre ces systèmes indépendants du genre du locuteur. La première solution est un mélange de deux modèles génératifs PLDA dépendants du genre. En effet, une simple reformulation du logarithme du rapport de vraisemblance (LLR) du modèle de base selon les lois de la probabilité permet l'obtention de ce nouveau modèle de mélange (PLDA-M) indépendant du genre. La deuxième solution, quant à elle, est proposée dans le contexte du classificateur à base de la similarité angulaire du cosinus (SAC). L'idée est de combiner les scores calculés par deux SAC dépendantes du genre à l'aide d'une somme pondérée dont les facteurs sont obtenus à partir d'un détecteur de genre. Sans aucune information sur le genre des locuteurs, cette combinaison (SAC-C) a permis l'obtention des résultats semblables à ceux obtenus par les SAC dépendantes du genre. De plus les résultats de la SAC-C (« *det*5 » : EER des

hommes est de 1,67 % et celui des femmes est de 2,62 %) sont comparables à ceux obtenus par le PLDA-M (« *det*5 » : EER des hommes est de 1,81 % et celui des femmes est de 2,46%) dans le contexte de la parole téléphonique. Or, les résultats du modèle génératif (PLDA-M) (« *det*2 » : EER des hommes est de 3,87 % et celui des femmes est de 2,03 %) restent généralement supérieurs à ceux de la SAC-C (« *det*2 » : EER des hommes est de 4,51 % et celui des femmes est de 2,23 %) lorsque testés vis-à-vis de la parole microphonique.

Dans la deuxième partie de cette thèse, nous avons présenté une étude détaillée de l'application d'une nouvelle version de l'algorithme non paramétrique de *Décalage de la moyenne* (MS) au problème du regroupement en locuteurs des segments vocaux non étiquetés. Cette nouvelle version de *Décalage de la moyenne* que nous avons proposée est basée sur la distance angulaire de cosinus. En effet, le MS à base de la distance angulaire du cosinus peut être vu comme une adaptation réussie de la version de base de *Décalage de la moyenne* à notre problème de regroupement en locuteurs des segments représentés dans l'espace des i-vecteurs. De plus, nous avons testé deux mécanismes exploitant cette nouvelle version de MS afin d'accomplir la tâche du regroupement, soit, la *stratégie totale* du regroupement (STR) et la *stratégie sélective* du regroupement (SSR). Nous avons d'abord démontré l'efficacité de l'algorithme de MS à base de la distance angulaire du cosinus (un point d'égalité des impuretés égal à 9 %) par rapport à la version de base de MS (un point d'égalité des impuretés égal à 12,3 %) lorsque testés vis-à-vis du problème de regroupement en locuteurs des grandes bases de données (Senoussaoui, *et al.*, 2013b). De même, nous avons prouvé comment cette simple méthode (c.-à-d. le MS à base de la distance angulaire du cosinus) est capable de gérer plusieurs problèmes difficiles reliés à la tâche de la structuration en tours de paroles, à savoir, la courte durée des tours de parole, le nombre inconnu et variable des locuteurs et les durées variables des conversations à structurer. Finalement, la preuve mathématique de convergence justifiant notre extension de l'algorithme de base de *Décalage de la moyenne* au nouvel algorithme à base de la distance angulaire du cosinus est fournie dans l'Annexe I de cette thèse.

Pour récapituler, la première partie de cette thèse montre comment nous avons conçu les premiers systèmes de vérification du locuteur qui sont à la fois indépendants du type du canal

de transmission (téléphone/microphone) et du genre du locuteur (homme/femme). Ces systèmes sont capables d'atteindre les résultats de l'état de l'art sans profiter des informations concernant le type du canal et le genre des locuteurs. Dans la deuxième partie, nous avons présenté un système de regroupement en locuteurs à base d'une nouvelle version de l'élégant algorithme de *Décalage de la moyenne*. Avec une seule passe du regroupement (c.-à-d. sans la re-segmentation de Viterbi), nous avons réussi à obtenir une amélioration relative de 37 % (de 19,8 % à 12,4 %) telle que mesurée par le taux d'erreur de la segmentation DER par rapport aux résultats présentés dans (Vaquero, 2011). De plus, nos résultats sont également comparables à ceux de l'état de l'art, tels que rapportés par d'autres auteurs (Dalmasso, *et al.*, 2005)(Castaldo *et al.*, 2008)(Shum, *et al.*, 2012).

❖ **Travaux futurs**

Bien que nos systèmes présentés dans la première partie de cette thèse soient capables d'atteindre les résultats de l'état de l'art en terme des métriques MinDCF et EER sans aucun besoin de l'information concernant le type du canal et du genre du locuteur, le choix du bon seuil de décision pour ces systèmes reste toujours dépendant du type du canal. Une recherche dans le domaine de la calibration des scores (Brummer, *et al.*, 2006) serait sans doute bénéfique afin d'accroitre davantage la robustesse des systèmes de la vérification du locuteur.

En ce qui concerne le problème de la structuration en tours de parole, nous avons constaté l'efficacité du réajustement des frontières entre les tours de parole via l'algorithme de Viterbi. L'utilisation de la nouvelle segmentation obtenue de cette manière pourrait servir comme une bonne initialisation d'une deuxième passe d'algorithme de *Décalage de la moyenne*. Une complication intéressante qui se poserait lorsqu'on explore cette piste, est que les tours de parole seraient d'une durée beaucoup plus variable que lors de la première passe où nous avons adopté une segmentation uniforme. L'incertitude de l'estimation d'un i-vecteur dépend directement de la durée du segment vocal, nous suggérons ainsi la prise en compte de cette incertitude tout comme dans (Kenny, *et al.*, 2013).

Somme toute, ma contribution la plus pertinente est d'abord le fait de rendre les deux systèmes de l'état de l'art de la vérification du locuteur indépendants du type du canal et du genre du locuteur sans aucune perte de performances. De plus, nous avons proposé une nouvelle version de l'algorithme de *Décalage de la moyenne* à base de la distance angulaire capable d'atteindre les résultats de l'art de la structuration en tours de parole (*Diarization*).

ANNEXE I

PREUVE MATHÉMATIQUE DE CONVERGENCE DE L'ALGORITHME DE DÉCALAGE DE LA MOYENNE À BASE DE LA DISTANCE ANGULAIRE DU COSINUS

Dans cette annexe, nous présentons la preuve mathématique de convergence de la nouvelle version de l'algorithme de *Décalage de la moyenne* à base de la distance angulaire du cosinus présenté dans cette thèse. Cette preuve est très similaire à celle du théorème 1 présentée dans cet article (Comaniciu, *et al.*, 2002).

Théorème 1 de (Comaniciu, *et al.*, 2002) stipule : « *si le noyau k a un profil convexe et monotone décroissant, la séquence des positions* $\{\hat{f}_i\}_{i=1,2...}$ *converge, et est monotone croissante* ».

Supposons d'abord que tous les vecteurs de notre ensemble de données sont contraints d'être projetés dans l'hypersphère unitaire, et ce, en normalisant à **1** leurs normes euclidiennes durant tout le processus de convergence de l'algorithme de MS.

$$\hat{f}_{j+1} - \hat{f}_j = c\sum_{j=1}^{n}\left[k\left(\frac{1-\mathbf{y}_{i+1}\cdot\mathbf{x}_j}{h}\right) - k\left(\frac{1-\mathbf{y}_i\cdot\mathbf{x}_j}{h}\right)\right].$$

Dû à la convexité de notre profil :

$$k(x_2) - k(x_1) \geq k'(x_1)(x_2 - x_1)$$

et étant donné que $g(x) = -k(x)$ (voir équation 6.5) :

$$k(x_2) - k(x_1) \geq g(x_1)(x_1 - x_2)$$

nous obtenons alors :

$$\hat{f}_{i+1} - \hat{f}_i \geq c\sum_{j=1}^{n} g\left(\frac{1-\mathbf{y}_i \cdot \mathbf{x}_j}{h}\right)\left[\frac{\left(1-\mathbf{y}_i \cdot \mathbf{x}_j\right)-\left(1-\mathbf{y}_{i+1} \cdot \mathbf{x}_j\right)}{h}\right]$$
$$= c\sum_{j=1}^{n} g\left(\frac{1-\mathbf{y}_i \cdot \mathbf{x}_j}{h}\right)\mathbf{x}_j\left(\frac{\mathbf{y}_{i+1}-\mathbf{y}_i}{h}\right)$$

À partir des équations (6.7) et (6.10), nous savons déjà que la position ($\mathbf{y}_{i+1}$) est égale au vecteur moyen pondéré de toutes les données se situant à l'intérieur du noyau, alors :

$$\sum_{j=1}^{n} g\left(\frac{1-\mathbf{y}_i \cdot \mathbf{x}_j}{h}\right)\mathbf{y}_{i+1} = \sum_{j=1}^{n} g\left(\frac{1-\mathbf{y}_i \cdot \mathbf{x}_j}{h}\right)\mathbf{x}_j \ .$$

De plus,

$$\hat{f}_{i+1} - \hat{f}_i \geq c\sum_{j=1}^{n} g\left(\frac{1-\mathbf{y}_i \cdot \mathbf{x}_j}{h}\right)\mathbf{y}_{i+1}\left(\frac{\mathbf{y}_{i+1}-\mathbf{y}_i}{h}\right)$$
$$= c\sum_{j=1}^{n} g\left(\frac{1-\mathbf{y}_i \cdot \mathbf{x}_j}{h}\right)\left(\frac{1-\mathbf{y}_{i+1} \cdot \mathbf{y}_i}{h}\right) \geq 0$$

avec une égalité, si et seulement si, $\mathbf{y}_{i+1} = \mathbf{y}_i$.

La séquence $\{\hat{f}_i\}_{i=1,2,\ldots}$ est délimitée et monotone croissante, donc elle est convergente. À vrai dire, cet argument ne démontre pas que la séquence $\{\mathbf{y}_i\}_{i=1,2,\ldots}$ soit forcément convergente, car il est possible d'établir des exemples pathologiques avec lesquels la séquence $\{\hat{f}_i\}_{i=1,2,\ldots}$ converge tandis que la séquence $\{\mathbf{y}_i\}_{i=1,2,\ldots}$ ne converge pas. Or, le présent argument démontre la convergence de l'algorithme de *Décalage de la moyenne* dans le même sens que la convergence de l'algorithme de EM est démontrée dans l'article (Dempster, *et al.*, 1977).

ANNEXE II

REPRÉSENTATIONS GRAPHIQUES DES EFFETS DES DIFFÉRENTES ÉTAPES DE LA NORMALISATION DES I-VECTEURS DANS LE CONTEXTE DE L'ALGORITHME DE DÉCALAGE DE LA MOYENNE

Dans cette deuxième annexe, nous présentons quelques illustrations graphiques des deux premières composantes des i-vecteurs. Chaque illustration montre les i-vecteurs représentant les segments d'une conversation donnée du corpus de test des données CallHome. Chaque figure contient quatre sous-illustrations montrant les effets de l'application successive des méthodes de la normalisation (BCCN, normalisation à 1 de la norme euclidienne -Norme Euc. 1-, PCA, deuxième Norme Euc. 1) sur les i-vecteurs de la même conversation. Les i-vecteurs représentés dans les quatre sous-illustrations sont tous colorés en fonction des mêmes étiquettes des groupes détectés par une seule passe de la version d'algorithme de MS à base de la distance angulaire du cosinus.

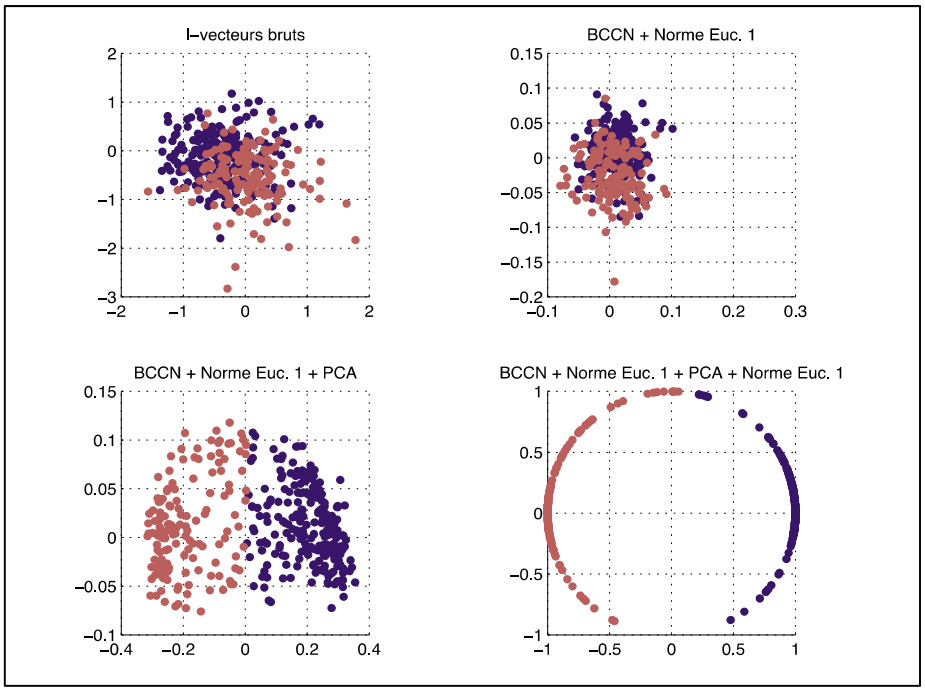

Figure II.1 Effets de la normalisation des i-vecteurs de la conversation *iahb*

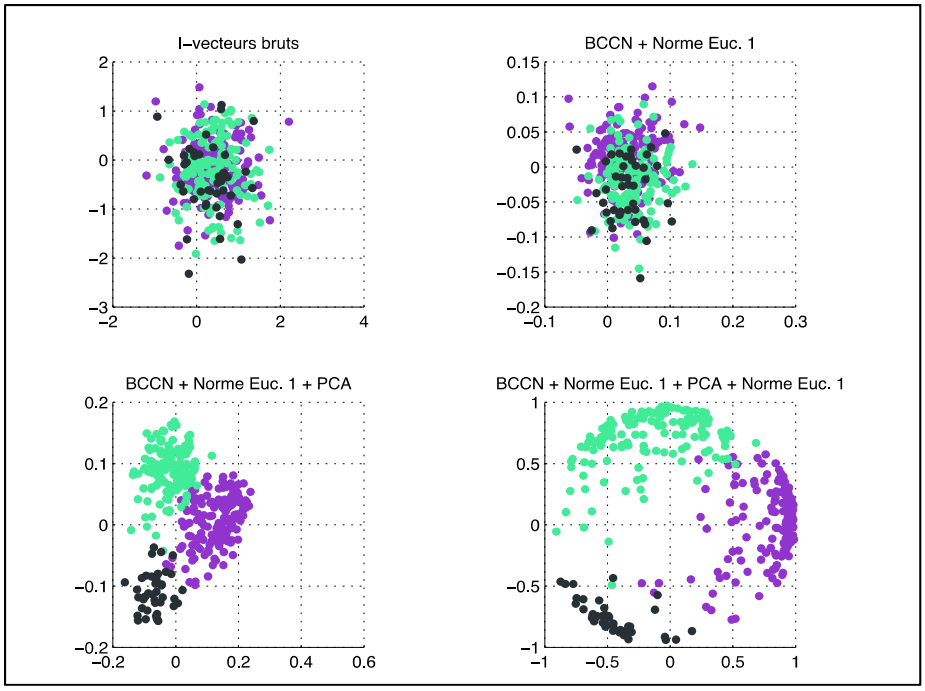

Figure II.2 Effets de la normalisation des i-vecteurs de la conversation *iabi*

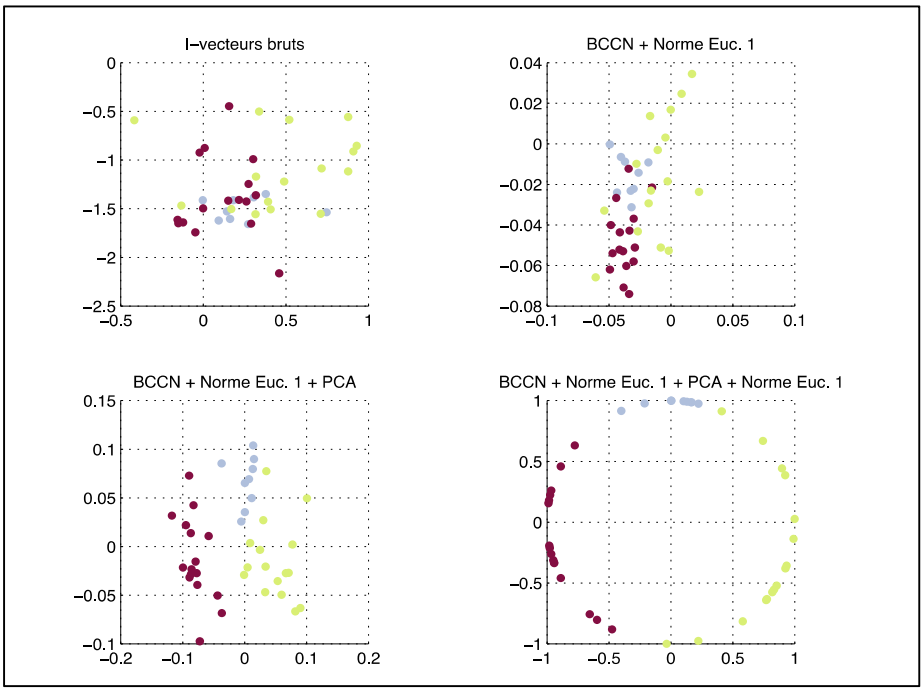

Figure II.3 Effets de la normalisation des i-vecteurs de la conversation *iaab*

ANNEXE III

INTERVALLES DE CONFIANCE CONCERNANT LES RÉSULTATS DE LA VÉRIFICATION

Dans cette dernière annexe, nous fournirons des informations supplémentaires concernant l'incertitude de l'estimation des mesures d'évaluation des performances de la vérification du locuteur proposées par NIST, à savoir, l'EER et les minimums DCF (c.-à-d. DCF_08 et DCF_10 fournies respectivement durant les campagnes d'évaluation de NIST SRE 2008 et SRE 2010) (voir Section 1.2.7). Dans le cadre de ce travail, nous avons calculé des intervalles de confiance au moyen de la procédure suivante :

1. Diviser d'une manière aléatoire la liste des essais de NIST en *L* sous-listes.
2. Évaluer séparément l'EER et les minimums DCF pour chaque sous-liste.
3. Refaire les étapes 1 et 2 *N* fois, tout en accumulant les *N* x *L* résultats calculés (c.-à-d. l'EER et les minimums DCF).
4. Évaluer la moyenne ainsi que l'écart-type pour chaque mesure (c.-à-d. l'EER et les minimums DCF) en utilisant ces *N* x *L* résultats.

Étant donné que, dans le cadre général de cette thèse, l'utilisation de ces informations concernant l'incertitude des mesures d'évaluation n'était certainement pas une préoccupation, nous avons adopté cette simple procédure détaillée ci-dessus (où N=10 et L=100). Cependant, le lecteur intéressé peut se référer à (Wu, et al., 2011) pour avoir une étude détaillée de ce sujet.

Tous les résultats obtenus (c.-à-d. la moyenne et l'écart-type) à partir des mesures de performances des deux systèmes adoptés dans cette thèse, soit les systèmes PLDA et la SAC (dont les résultats sont présentés au chapitre 5) seront regroupés dans les tableaux de cette annexe. Les résultats obtenus sur les listes complètes de NIST sont également fournis dans les colonnes grises (nommées **Tot.**) de ces tableaux.

Tableau III.1 Résultats (avec intervalles de confiance) des différents systèmes PLDA testés sur la liste det5 (téléphone/téléphone) de NIST SRE 2010.

		EER (%)			MinDCF_08			MinDCF_10		
		Tot.	μ	σ	**Tot.**	μ	σ	**Tot.**	μ	σ
Femmes	PLDA-DG	**2.47**	2.40	0.27	**0.124**	0.118	0.008	**0.387**	0.373	0.033
	PLDA-IG	2.75	2.62	0.27	0.133	0.126	0.008	0.415	0.397	0.036
	PLDA-M	**2.46**	2.40	0.27	**0.124**	0.118	0.008	**0.388**	0.374	0.034
Hommes	PLDA-DG	**1.81**	1.71	0.20	**0.096**	0.091	0.008	**0.320**	0.296	0.030
	PLDA-IG	2.00	1.93	0.21	0.112	0.104	0.008	0.386	0.362	0.032
	PLDA-M	**1.81**	1.71	0.20	**0.096**	0.091	0.008	**0.322**	0.297	0.031

Tableau III.2 Résultats (avec intervalles de confiance) des différents systèmes PLDA testés sur la liste det2 (interview/interview) de NIST SRE 2010.

		EER (%)			MinDCF_08			MinDCF_10		
		Tot.	μ	σ	**Tot.**	μ	σ	**Tot.**	μ	σ
Femmes	PLDA-DG	3.86	3.80	0.23	0.190	0182	0.007	0.543	0.536	0.015
	PLDA-IG	**3.80**	3.77	0.22	**0.187**	0.179	0.006	**0.536**	0.530	0.016
	PLDA-M	3.87	3.81	0.20	0.190	0.182	0.007	0.541	0.533	0.015
Hommes	PLDA-DG	**2.02**	1.97	0.18	**0.097**	0.092	0.005	**0.363**	0.358	0.017
	PLDA-IG	2.11	2.03	0.19	0.098	0.093	0.005	0.397	0.390	0.019
	PLDA-M	**2.03**	1.98	0.17	**0.097**	0.092	0.005	**0.365**	0.360	0.018

Tableau III.3 Résultats (avec intervalles de confiance) des différents systèmes PLDA testés sur les listes det1, det3 et det4, regroupant les locuteurs hommes et femmes, de NIST SRE 2010.

		EER (%)			MinDCF_08			MinDCF_10		
		Tot.	μ	σ	**Tot.**	μ	σ	**Tot.**	μ	σ
*det*1	PLDA-DG	1.58	1.34	0.24	0.070	0.064	0.009	0.246	0.237	0.028
	PLDA-IG	**1.44**	1.36	0.27	0.071	0.066	0.009	0.262	0.250	0.028
	PLDA-M	1.58	1.34	0.24	0.070	0.064	0.009	0.246	0.237	0.030
*det*3	PLDA-DG	2.68	2.55	0.39	0.126	0.117	0.012	0.402	0.385	0.037
	PLDA-IG	**2.57**	2.44	0.39	**0.124**	0.116	0.013	0.439	0.417	0.037
	PLDA-M	2.68	2.55	0.41	0.125	0.116	0.012	0.397	0.381	0.036
*det*4	PLDA-DG	2.90	2.74	0.44	0.128	0.121	0.013	0.385	0.369	0.033
	PLDA-IG	3.05	2.91	0.48	0.133	0.124	0.013	0.403	0.385	0.032
	PLDA-M	2.90	2.74	0.45	0.129	0.121	0.013	0.384	0.368	0.032

Tableau III.4 Résultats (avec intervalles de confiance) des différents systèmes à base de la SAC testés sur la liste det5 (téléphone/téléphone) de NIST SRE 2010. Les résultats du PLDA-M sont repris à des fins de comparaison.

			EER (%)			MinDCF_08			MinDCF_10		
			Tot.	μ	σ	**Tot.**	μ	σ	**Tot.**	μ	σ
Femmes	SAC-DG	*zt-norm*	**2.62**	2.53	0.24	0.151	0.141	0.008	0.551	0.524	0.036
		s-norm	2.75	2.63	0.28	**0.136**	0.128	0.009	0.452	0.427	0.035
	SAC-IG	*zt-norm*	3.82	3.78	0.25	0.258	0.242	0.012	0.698	0.673	0.029
		s-norm	3.07	2.98	0.26	0.158	0.148	0.009	0.517	0.498	0.040
	SAC-C	*zt-norm*	**2.64**	2.53	0.24	0.150	0.140	0.009	0.550	0.524	0.034
		s-norm	2.73	2.63	0.28	**0.135**	0.127	0.009	**0.448**	0.426	0.036
Hommes	SAC-DG	*zt-norm*	**1.67**	1.62	0.20	**0.091**	0.087	0.007	0.406	0.375	0.047
		s-norm	1.99	1.84	0.25	0.102	0.096	0.008	**0.364**	0.338	0.037
	SAC-IG	*zt-norm*	2.52	2.45	0.23	0.164	0.153	0.009	0.704	0.672	0.039
		s-norm	2.25	2.18	0.23	0.133	0.124	0.008	0.580	0.547	0.043
	SAC-C	*zt-norm*	**1.67**	1.62	0.20	**0.091**	0.087	0.000	0.411	0.378	0.047
		s-norm	1.98	1.84	0.25	0.102	0.096	0.008	**0.368**	0.341	0.037

Tableau III.5 Résultats (avec intervalles de confiance) des différents systèmes à base de la SAC testés sur la liste det2 (téléphone/téléphone) de NIST SRE 2010. Les résultats du PLDA-M sont repris à des fins de comparaison.

			EER (%)			MinDCF_08			MinDCF_10		
			Tot.	μ	σ	**Tot.**	μ	σ	**Tot.**	μ	σ
Femmes	SAC-DG	*zt-norm*	4.53	4.47	0.23	0.236	0.226	0.007	0.700	0.693	0.017
		s-norm	4.73	4.62	0.22	**0.226**	0.217	0.007	**0.624**	0.616	0.018
	SAC-IG	*zt-norm*	7.40	7.33	0.23	0.443	0.427	0.009	0.840	0.835	0.014
		s-norm	4.88	4.81	0.22	0.256	0.245	0.008	0.712	0.705	0.017
	SAC-C	*zt-norm*	**4.51**	4.46	0.24	0.236	0.226	0.007	0.695	0.689	0.018
		s-norm	4.70	4.61	0.25	**0.225**	0.217	0.007	**0.622**	0.613	0.017
Hommes	SAC-DG	*zt-norm*	**2.23**	2.19	0.18	0.117	0.111	0.005	0.493	0.486	0.020
		s-norm	2.38	2.34	0.17	**0.114**	0.109	0.005	**0.414**	0.407	0.019
	SAC-IG	*zt-norm*	3.80	3.72	0.19	0.254	0.242	0.008	0.891	0.881	0.015
		s-norm	2.69	2.63	0.18	0.150	0.143	0.006	0.576	0.566	0.018
	SAC-C	*zt-norm*	2.25	2.19	0.18	0.117	0.112	0.006	0.494	0.487	0.020
		s-norm	2.39	2.35	0.18	**0.115**	0.110	0.006	**0.414**	0.408	0.018

Tableau III.6 Résultats (avec intervalles de confiance) des différents systèmes à base de la SAC testés sur les listes det1, det3 et det4 de NIST SRE 2010, regroupant les locuteurs hommes et femmes.

			EER (%)			MinDCF_08			MinDCF_10		
			Tot.	μ	σ	**Tot.**	μ	σ	**Tot.**	μ	σ
*det*1	SAC-DG	*zt-norm*	1.89	3.45	0.22	0.110	0.175	0.007	0.514	0.625	0.021
		s-norm	**1.62**	3.62	0.24	**0.080**	0.168	0.007	**0.328**	0.527	0.019
	SAC-IG	*zt-norm*	5.15	5.97	0.21	0.385	0.396	0.010	0.856	0.895	0.012
		s-norm	2.04	3.88	0.22	0.122	0.199	0.008	0.552	0.689	0.020
	SAC-C	*zt-norm*	1.88	3.45	0.22	0.110	0.175	0.007	0.515	0.623	0.020
		s-norm	**1.62**	3.63	0.23	**0.080**	0.168	0.007	**0.328**	0.525	0.020
*det*3	SAC-DG	*zt-norm*	**3.00**	2.88	0.42	**0.138**	0.129	0.012	0.502	0.482	0.041
		s-norm	3.28	3.14	0.47	0.146	0.137	0.014	0.487	0.467	0.038
	SAC-IG	*zt-norm*	3.90	3.77	0.36	0.256	0.239	0.015	0.848	0.831	0.034
		s-norm	3.25	3.10	0.44	0.164	0.153	0.013	0.639	0.614	0.042
	SAC-C	*zt-norm*	**3.00**	2.89	0.42	**0.138**	0.128	0.013	0.491	0.471	0.042
		s-norm	3.29	3.15	0.44	0.146	0.136	0.013	**0.480**	0.460	0.040
*det*4	SAC-DG	*zt-norm*	3.68	3.50	0.54	0.161	0.151	0.014	0.584	0.562	0.037
		s-norm	**3.65**	3.44	0.54	**0.147**	0.138	0.014	**0.434**	0.417	0.035
	SAC-IG	*zt-norm*	5.58	5.42	0.46	0.362	0.338	0.018	0.889	0.877	0.022
		s-norm	3.86	3.68	0.52	0.177	0.165	0.014	0.608	0.588	0.041
	SAC-C	*zt-norm*	3.68	3.49	0.55	0.161	0.151	0.015	0.585	0.563	0.038
		s-norm	**3.65**	3.44	0.55	**0.147**	0.138	0.014	**0.436**	0.418	0.033

BIBLIOGRAPHIE

Bellman, R. E. (1957). « *Dynamic programming*. » Princeton University Press. ISBN 978-0-691-07951-6. Republié: Richard Ernest Bellman (2003). « *Dynamic Programming*. » Courier Dover Publications. ISBN 978-0-486-42809-3.

Bengio Y. et LeCun, Y. (2007). « Scaling learning algorithms toward AI. » in Large Scale Kernel Machine. Cambridge, MA: MIT Press, pp. 321–360.

Bengio, Y. (2009). « Learning deep architectures for AI. » Foundations and Trends in Machine Learning, vol.2, pp. 1-127.

BenZeghiba, M. et Bourlard, H. (2003). « On the Combination of Speech and Speaker Recognition. » In: Proc. Eighth European Conf. on Speech Communication and Technology, Geneva, Switzerland, pp. 1361–1364.

Bimbot, F. Bonastre, J-F. Fredouille, C. Gravier, G. Magrin-Chagnolleau, I. Meignier, S. Merlin, T. Ortega-Garcia, J. Petrovska-Delacrétaz, D. et Reynolds, D. (2004). « A Tutorial on Text-Independent Speaker Verification. » EURASIP Journal on Applied Signal Processing 2004:4, 430–451.

Bishop CM. (2007). « Pattern Recognition and Machine Learning. » Information Science and Statistics. 1st ed. Springer.

Bousquet, P. larcher, A. Matrouf, D. Bonastre, J. et Plchot, O. (2012). « Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis. » Proc. The Speaker and Language Recognition Workshop. Singapur: International Speech Communication Association, pp. 157-164.

Brummer N. et du Preez, J. (2006). « Application-independent Evaluation of Speaker Detection. » Comput. Speech Lang. (UK), 20(2-3) :230 – 75.

Brummer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, R, Karaat, M., Leeuwen, D. v., Matejka, P., Schwarz, P., et Strasheim, A. (2007). « Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006.» Audio, Speech, and Language Processing, IEEE Transactions on , vol.15, no.7, pp.2072,2084.

Burget, L. Plchot, O. Cumani, S. Glembek, O. Matejka, P. Brummer, N. (2011). « Discriminatively trained Probabilistic Linear Discriminant Analysis for speaker verification. » Acoustics, Speech and Signal Processing (ICASSP), pp.4832-4835.

Campbell, J.P. (1997). « Speaker Recognition : a tutorial. » Proc. IEEE (USA), 85(9) :1437 – 62.

Campbell, W., Sturim, D., Navratil, J., Shen, W., et Reynolds, D. (2007). « The MIT-LL/IBM 2006 Speaker Recognition System: High-Performance Reduced-Complexity Recognition.» Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol.4, no., pp.IV-217,IV-220, 15-20.

Campbell, W.M. Campbell, J.P. Reynolds, D.A. Singer, E. et Torres-Carrasquillo, P.A. (2006). « Support vector machines for speaker and language recognition. » Comput. Speech Lang. (UK), 20(2-3) : 210 – 29.

Castaldo, F. Colibro, D. Dalmasso, E. Laface, P. et Vair, C. (2008). « Stream-based Speaker Segmentation Using Speaker Factors and Eigenvoices. » in Proceedings of ICASSP.

Chen, S. S. et Gopalakrishnan, P. (1998). « Clustering via the Bayesian Information Criterion with Applications in Speech Recognition. » in ICASSP'98, vol. 2, Seattle, USA, pp. 645–648.

Cheng, Y. (1995). « Mean Shift, Mode Seeking, and Clustering. » IEEE Trans. PAMI, vol. 17, no. 8, pp. 790-799.

Comaniciu, D. et Meer, P. (2002). « Mean Shift: A Robust Approach Toward Feature Space Analysis. » IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603 – 619.

Comaniciu, D. Ramesh, V. et Meer, P. (2001). « The Variable Bandwidth Mean Shift and Data-Driven Scale Selection. » Proc. Eighth Conference on Computer Vision, vol. I, pp. 438-445.

Dahl, G. et Hinton, G. (2010). « Phone recognition with the mean-covariance restricted Boltzmann machine. » in Advances in Neural Information Processing.

Dalmasso, E. Laface, P. Colibro, D. et Vair, C. (2005). « Unsupervised Segmentation and Verification of Multi-Speaker Conversational Speech. » Proc. Interspeech.

Davis, S. et Mermelstein, P. (1980). « Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. » IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4) :357–366.

Dehak, N. Dehak, R. Glass, J. Reynolds, D. et Kenny, P. (2010). « Cosine Similarity Scoring without Score Normalization Techniques. » Proc. IEEE Odyssey Workshop, Brno, Czech Republic.

Dehak, N. Dehak, R. Kenny, P. Brummer, N. Ouellet, P. et Dumouchel, P. (2009). « Support Vector Machines Versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. » Proc. Interspeech, pp. 1559-1562, Brighton, UK.

Dehak, N. Karam, Z. Reynolds, D. Dehak, R. Campbell, W. et Glass, J. (2011a). « A Channel-Blind System for Speaker Verification. » Proc. ICASSP, pp. 4536-4539, Prague, Czech Republic.

Dehak, N. Kenny, P. Dehak, R. Dumouchel, P. et Ouellet, P. (2011b). « Front-end Factor Analysis for Speaker Verification. » IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, pp. 788-798.

Dehak, N. Torres-Carrasquillo, P. Reynolds, D. et Dehak, R. (2011c). « Language Recognition via Ivectors and Dimensionality Reduction. » Proc. Interspeech, pp. 857-860, Florence, Italy.

Dempster, A. P. Laird, N. M. et Rubin, D. B. (1977). « Maximum Likelihood from Incomplete Data via the EM Algorithm. » Journal of the Royal Statistical Society, Series B (Methodological), vol. 39, no. 1, pp. 1–38.

Deng, L. Hinton, G. et Kingsbury, B. (2013). « New types of Deep Neural Network Learning for speech recognition and related applications: An overview. » Proceedings ICASSP 2013, pp. 8599-8603.

Doddington, G. (2001). « Speaker Recognition Based on Idiolectal Differences Between Speakers. » In Proc. of EUROSPEECH, pages 2521–2524, Scandinavia.

Farrell, K.R. Mammone, R.J. et Assaleh. K.T. (1994). « Speaker Recognition Using Neural Networks and Conventional Classifiers. » IEEE Trans. Speech Audio Process. (USA), 2(1) :194 – 205.

Ferrer, L., Shriberg, E., Kajarekar, S., Stolcke, A., Sonmez, K., Venkataraman, A., et Bratt, H. (2006). « The Contribution of Cepstral and Stylistic Features to SRI's 2005 NIST Speaker Recognition Evaluation System. », Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol.1, no., pp.I,I,

Fisher, R. (1936). « The use of multiple measurements in taxonomic problems » Annals Eugen., vol. 7, pp. 179–188.

Fukunaga, K. et Hostetler, L. (1975). « The Estimation of The Gradient of A Density Function, With Applications In Pattern Recognition. » IEEE Transaction on Information Theory, vol. 21, no. 1, pp. 32–40.

Furui, S. (1981) « Cepstral Analysis Technique for Automatic Speaker Verification. » Acoustics, Speech and Signal Processing, IEEE Transactions on , vol.29, no.2, pp.254,272.

Furui, S. (2005). « 50 Years of Progress in Speech and Speaker Recognition Research. » ECTI Transactions On Computer And Information Technology, vol.1, no.2.

Furui, S. (2008). « Speaker recognition. » Scholarpedia, vol.3, pp 3715.

Garcia-Romero, D. (2011). « Analysis of I-Vector Length Normalization in Speaker Recognition Systems. » Proceedings of Interspeech, Florence, Italy.

Gauvain, J.-L. et Lee, C.-H. (1994). « Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains » Speech and Audio Processing, IEEE Transactions on, vol. 2, no. 2, pp. 291–298.

Gupta, V. Kenny, P. Ouellet, P. Boulianne, G. et Dumouchel, P. (2007). « Combining Gaussianized/non-Gaussianized Features to Improve Speaker Diarization of Telephone Conversations, » IEEE Signal Processing Letters, 14 (12), pp. 1040-1043.

Hatch, A. Kajarekar, S. et Stolcke, A. (2006). « Within- class covariance normalization for svm-based speaker recognition. » INTERSPEECH, 9th International Conference on Spoken Language Processing, vol. 3, pp. 1471 – 1474.

Hinton, G. Deng, L. Yu, D. Dahl, G. Mohamed, A. Jaitly, N. Senior, A. Vanhoucke, V. Nguyen, P. Sainath, T. et Kingsbury, B. (2012). « Deep Neural Networks for acoustic modeling in speech recognition. » IEEE Signal Processing Magazine, vol. 28, no. 6, pp. 82-97.

Hinton, G. E. et Sejnowski, T. J. (1983). « Optimal Perceptual Inference. » Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 448-453 Washington DC.

Hinton, G. E. Osindero, S. et Teh, Y. W. (2006). « A fast learning algorithm for deep belief nets. » Neural Computation, 18(7), pp. 1527-1554.

Jin, Q. Schultz, T. (2008). « Robust Far-Field Speaker Recognition under Mismatched Conditions. » Proc. 9th Annual Conference of the International Speech Communication Association.

Kajarekar, S. (2005). « Four Weightings and a Fusion: A Cepstral-SVM System for Speaker Recognition. » Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on, vol., no., pp.17,22, 27-27.

Kanagasundaram, A. Dean, D. Gonzalez-Dominguez, J. Sridharan, S. Ramos, D. et Gonzalez-Rodriguez, J. (2013). « Improving the PLDA based Speaker Verification in Limited Microphone Data Conditions. » INTERSPEECH, page 3674-3678. ISCA.

Kenny, P. (2010a). « Bayesian Speaker Verification with Heavy-Tailed Priors» Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic.

Kenny, P. (2012). « A Small Footprint i-Vector Extractor » Proc. Odyssey Speaker and Language Recognition Workshop, Singapore.

Kenny, P. Boulianne, G. et Dumouchel, P. (2005). « Eigenvoice Modeling With Sparse Training Data. » IEEE Trans. Speech Audio Processing, vol. 13, no. 3, pp. 345–359.

Kenny, P. Ouellet, P. Dehak, N. Gupta, V. et Dumouchel, P. (2008). « Study of Inter-Speaker Variability in Speaker Verification. » IEEE Transactions on Audio, Speech and Language Processing.

Kenny, P. Reynolds, D. et Castaldo, F. (2010b). « Diarization of Telephone Conversations using Factor Analysis. » IEEE Journal of Selected Topics in Signal Processing.

Kenny, P. Stafylakis, T. Ouellet, P. Alam, J. et Dumouchel, P. (2013). « PLDA for Speaker Verification with Utterances of Arbitrary Duration. » Proc. ICASSP, Vancouver, Canada.

Kinnunen, T. et Li, H. (2010). « An Overview of Text-Independent Speaker Recognition : from Features to Supervectors. » Speech Communication, 52(1) :12 – 40.

Kotti, M. Moschou, V. et Kotropoulos, C. (2008). « Speaker Segmentation and Clustering. » Signal Processing, Volume 88, Pages 1091-1124.

Li, K.-P. et Porter, J.E. (1988). « Normalizations and Selection of Speech Segments for Speaker Recognition Scoring » ICASSP-88., 1988 International Conference on Acoustics, Speech, and Signal Processing, 1988, 1 :595–598.

Martin A. et Przybocki, M. (2001). « Speaker Recognition in a Multi-Speaker Environment. » in Proceedings of Eurospeech.

Moattar, M. H. et Homayounpour, M. M. (2012). « A Review on Speaker Diarization Systems and Approaches. » Speech Commun. 54, 1065-1103.

Mohamed, A. Sainath, T. Dahl, G. Ramabhadran, B. Hinton, G. et Picheny, M. (2011). « Deep belief networks using discriminative features for phone recognition. » in Proc. ICASSP 2011, Zurich, Switzerland.

Pelecanos, J. et Sridharan, S. (2001). « Feature Warping for Robust Speaker Verification. » IEEE Odyssey : The Speaker and Language Workshope, pages 213-218, Crete, Greece.

Pigeon, S. Druyts, P. et Verlinde, P. (2000) « Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. » Digital Signal Processing.

Prince, S. J. D. et Elder, J. H. (2007). « Probabilistic Linear Discriminant Analysis for Inferences about Identity. » in Proc. 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, pp. 1–8.

Reynolds, D. A. (1992). « A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. » Ph.D. dissertation, Georgia Institute of Technology.

Reynolds, D. A. (1995) « Speaker Identification and Verification using Gaussian Mixture Speaker Models. » Speech Communication, Volume 17, Issues 1–2, Pages 91-108.

Reynolds, D. A. Dunm, R. B. et Laughlin, J. J. (2000a). « The Lincoln Speaker Recognition System : NIST EVAL2000. » Proceedings of International Conference on Spoken Language Processing.

Reynolds, D. A. Quatieri, T.-F. et Dunn, R.-B. (2000b). « Speaker Verification Using Adapted Gaussian Mixture Models. » Digital Signal Processing, Volume 10, Issues 1–3, Pages 19-41.

Salakhutdinov R. (2009). « Learning Deep Generative Models. » Ph.D. dissertation, University of Toronto.

Schwarz, G. (1978). « Estimating the Dimension of a Model. » The Annals of Statistic 6(2) :461– 464.

Senoussaoui, M. Dehak, N. Kenny, P. Dehak, R. et Dumouchel, P. (2012). « First Attempt at Boltzmann Machines for Speaker Recognition. » Proc. Odyssey Speaker and Language Recognition Workshop, Singapore.

Senoussaoui, M. Kenny, P. Brummer, N. de Villiers, E. et Dumouchel, P. (2011a). « Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition. » Proc. Interspeech 2011, Florence.

Senoussaoui, M. Kenny, P. Dehak, N. et Dumouchel, P. (2010). « An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech. » in Proc Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic.

Senoussaoui, M. Kenny, P. Dumouchel, P. et Castaldo, F. (2011b). « Well Calibrated Heavy Tailed Bayesian Speaker Verification for Microphone Speech. » Proc ICASSP, Prague, Czech Republic.

Senoussaoui, M. Kenny, P. Dumouchel, P. et Dehak, N. (2013a). « New Cosine Similarity Scorings to Implement Gender-Independent Speaker Verification. » Proc. Interspeech, Lyon, France.

Senoussaoui, M. Kenny, P. Dumouchel, P. et Stafylakis. T. (2013b). « Efficient Iterative Mean Shift Based Cosine Dissimilarity for Multi-Recording Speaker Clustering » Proc. ICASSP, Vancouver, Canada.

Senoussaoui, M. Kenny, P. Stafylakis, T. et Dumouchel, P. (2014). « A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization. » IEEE Trans. Audio, Speech and Language Proc., to appear, 2014.

Shum, S. Dehak, N. Chuangsuwanich, E. Reynolds, D. et Glass, J. (2011). « Exploiting Intra-Conversation Variability for Speaker Diarization. » Proc. Interspeech, pp. 945-948, Florence, Italy.

Shum, S. Dehak, N. Dehak, R. et Glass, J.R. (2013). « Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach. » IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 10, pp. 2015-2028.

Shum, S. Dehak, N. et Glass, J. (2012). « On the Use of Spectral and Iterative Methods for Speaker Diarization. » Proc. Interspeech, Portland, Oregon.

Smolensky, P. (1986). « Information processing in dynamical systems: Foundations of harmony theory. » Dans D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing, volume 1, chapter 6, pp. 194-281. MIT Press, Cambridge.

Soong, F. Rosenberg, A. Rabiner, L. et Juang, B. (1985). « A Vector Quantization Approach to Speaker Recognition. » Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’85, 10 :387–390.

Stafylakis, T. Katsouros, V. et Carayannis, G. (2010). « Speaker clustering via the mean shift algorithm. » The Speaker and Language Recognition Workshop - Odyssey-10, Brno, Czech Republic.

Stafylakis, T. Katsouros, V. Kenny, P. et Dumouchel, P. (2012). « Mean Shift Algorithm for Exponential Families with Applications to Speaker Clustering,” Proc. Odyssey Speaker and Language Recognition Workshop, Singapore.

Stafylakis, T. Kenny, P. Senoussaoui, M. and Dumouchel, P. (2012a). « Preliminary Investigation of Boltzmann Machine Classifiers for Speaker Recognition. » Proc. Odyssey Speaker and Language Recognition Workshop, Singapore.

Stafylakis, T. Kenny, P. Senoussaoui, M. and Dumouchel, P. (2012b). « PLDA Using Gaussian Restricted Boltzmann Machines with Application to Speaker Verification. » Proc. Interspeech, Portland, Oregon.

Sturim, D. Campbell, W. Dehak, N. Karam, Z. McCree, A. Reynolds, D. Richardson, F. Torres-Carrasquillo, P. Shum, S. (2011). « The MIT LL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition. » Proc. IEEE ICASSP, Prague, Czech Republic.

Sturim, D.E. Campbell, W.M. Reynolds, D.A. Dunn, R.B. Quatieri, T.F. (2007). « Robust Speaker Recognition with Cross-Channel Data: MIT-LL Results on the 2006 NIST SRE Auxiliary Microphone Task. » Proc. IEEE International Conference on Acoustics, Speech and Signal Processing., vol.4, no., pp.IV-49,IV-52.

Suh, J.-W. Sadjadi, S. O. Liu, G. Hasan, T. Godin, K. W. et Hansen, J. H.L. (2011). « Exploring hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA. » Proc. NIST Speaker Recognition Evaluation Workshop.

Tang, H. Chu, S.M. Hasegawa-Johnson M. et Huang, T.S. (2012). « Partially Supervised Speaker Clustering. » IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.34, no.5, pp. 959-971.

Tranter, S.E. et Reynolds, D.A. (2006) « An Overview of Automatic Speaker Diarization Systems. » Audio, Speech, and Language Processing, IEEE Transactions on, vol.14, no.5, pp.1557-1565.

Valente, F. (2005). « Variational Bayesian Methods for Audio Indexing. » Ph.D. dissertation, Eurecom.

Van Leeuwen, D. (2010). « Speaker Linking in Large Data Sets. » Proc. IEEE Odyssey Workshop, Brno, Czech Republic.

Van Leeuwen, D.-A. et Brummer, N. (2007). « An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. » in Speaker Classification, Lecture Notes in Computer Science / Artificial Intelligence, C. Muller, Ed. Heidelberg - New York - Berlin: Springer, vol. 4343.

Vapnik, V.N. (1998). « Statistical Learning Theory. » 1st ed. Wiley.

Vaquero Avilés-Casco C. (2011). « Robust Diarization For Speaker Characterization (Diarizacion Robusta Para Caracterizacion De Locutores) » dissertation de thèse, Université de Zaragoza, Espagne.

Vasilakakis, V. Cumani, S. et Laface, P. (2013). « Speaker recognition by means of Deep Belief Networks. » Biometric Technologies in Forensic Science, Nijmegen.

Wu, J. C. Martin, A. F. Greenberg, C. S. Kacker, R. N. (2011). « Uncertainties of measures in speaker recognition evaluation. » Proc. SPIE 8040, Active and Passive Signatures II.